The Good, the Bad and the Deadly

Characterisation of bacteria through DNA analysis

Cosentino, Salvatore; Larsen, Mette Voldby; Lund, Ole

Publication date:
2014

Document Version
Peer reviewed version

The Good, the Bad and the Deadly: characterisation of bacteria through DNA analysis

Salvatore Cosentino

November, 2013

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

Preface

This Ph.D. thesis was prepared at the Center for Biological Sequence Analysis (CBS), Department of System Biology, Technical University of Denmark (DTU), under the supervision of Associate Professor Mette Voldby Larsen and Professor Ole Lund.

The results presented in the thesis demonstrate how bacterial Whole Genome Sequecing (WGS) can be used by clinical microbiologists to characterise known and novel bacteria causing infections in human.

The thesis consists of an introduction, followed by a collection of four research manuscripts produced during my Ph.D. study at CBS and at the Center for Genomic Epidemiology (CGE) during the period 2010-2013

Salvatore Cosentino
Lyngby, November 2013

CONTENTS

Summary

Sixty years have passed from the discovery of the DNA's double helix structure, and at the time it would have been impossible to imagine that today we could receive our own DNA with information about our ancestors, or potential diseases we could get, based on the analysis of the DNA sequences. High-throughput DNA sequencing started a revolution, which is still ongoing, in biology and medicine, bringing new ways of studying diseases that could not have been possible before. Millions of people die every year from bacterial infections, and given the trends toward globalization of travel and commerce, with urbanization and an aging population, infectious diseases are coming back as a possible global threat. At present thousands of completely sequenced bacteria are publicly available, with many hundreds of genomes completed every year, also due to the constantly decreasing costs for sequencing a bacterial isolate. Using this invaluable data, scientists are bringing a revolution in the way clinical microbiology is done and in the way infectious diseases are diagnosed and treated.

My PhD project focused on the use of Bacterial Whole Genome Sequence (WGS) data for the prediction of pathogenic features, and the identification of bacterial species and subtypes, as well as phenotypic traits like antibiotic resistance. The thesis starts with an introduction to the problems related to infectious diseases, as well as the history of DNA, DNA sequencing and its application in clinical microbiology are overviewed. The thesis continues with three chapters and concludes with a discussion on possible future applications of DNA sequence data in clinical and public health microbiology.

Chapter one introduces the challenges related to the use of WGS for the prediction of bacterial pathogenicity features. The manuscript included in this chapter describes a prediction method for bacterial pathogenicity, which was also the main topic of my PhD studies.

Chapter two introduces the possibility to identify bacterial species and subtypes through DNA sequence analysis and includes two manuscripts about this topic.

In chapter three the problem of antibiotic resistance is discussed and a manuscript describing a method for identifying antimicrobial resistant genes from bacterial WGS data is included.

The thesis terminates with a discussion on the possible future applications of sequencing technologies in clinical and public health microbiology.

Dansk resumé

Der er gået 60 år siden DNAs dobbelthelix struktur blev identificeret i en tid, hvor det var umuligt at forestille sig, at vi i dag kan få adgang til vores egen DNA, inklusiv information om vores forfædre og hvilke sygdomme vi er prædisponeret for, baseret på analyser af DNA sekvensen. High-throughput DNA sekventering startede en fortløbende revolution indenfor biologien og den medicinske verden. Denne revolution har medført nye måder at studere sygdomme på, som ikke ville have været mulige tidligere. Millioner af mennesker dør hvert år som følge af bakterielle infektioner, og givet den nuværende trend mod øget globalisering med rejser og samhandel, medregnet den øgede urbanisering og en aldrende befolkning, er det sandsynligt at infektiøse sygdomme vil genopstå som en global trussel. Der er i dag tusindvis af fuldt sekventerede bakterier offentligt tilgængeligt, og hundreder af yderligere genomer færdiggøres hvert år - også p.g.a. den konstant faldende pris for sekventeringen af et bakterieisolat. Ved at anvende dette uvurderlige data, muliggør forskere en revolution i forhold til den måde klinisk mikrobiologi udføres og den måde hvorpå infektiøse sygdomme diagnostikeres og behandles.

Mit PhD projekt har hovedsageligt været fokuseret på anvendelsen af Hel Genom Sekvens (HGS) data til forudsigelse af patogene træk og identifikation af bakterieart og -stamme, såvel som fænotypiske træk som antibiotika resistens. Afhandlingen indledes med en introduktion til de problemer, der er relateret til infektiøse sygdomme, såvel som historien bag DNA, DNA sekventering og dens brug indenfor klinisk mikrobiologi. Afhandlingen fortsætter med tre kapitler og slutter af med en diskussion om de mulige fremtidige anvendelser af DNA sekvensdata i forbindelse med klinisk medicin og indenfor folkesundhed.

Kapitel et introducerer de udfordringer, der er forbundet med brug af HGS data til forudsigelse af bakterielle patogene træk. Manuskriptet inkluderet i dette kapitel beskriver en metode til forudsigelse af bakteriel patogenesitet, hvilket også har været fokuspunktet for mit PhD studie.

Kapitel to introducerer muligheden for at identificere bakterieart og stamme via DNA sekvensanalyser og inkluderer to artikler om dette emne.

I kapitel tre diskuteres problemet med antibiotika resistens og en metode til identifikation af antibiotika resistens gener fra bakteriel HGS data er inkluderet.

Afhandlingen afsluttes med en diskussion om de mulige fremtidige anvendelsesmuligheder for sekvensteknologi i forbindelse med klinisk mikrobiologi og indenfor folkesundhed.

Acknowledgements

It was a great pleasure to do my work as a PhD student at the Center for Biological Sequence Analysis (CBS), surrounded by exceptional people always ready for ready help to me in when I needed, to have stimulating discussions about scientific and non-scientific matters. I take here the chance to thank all CBSians. A big thank also goes to people at National Food Department and people at the Center for Genomic Epidemiology, from which my PhD studies were funded.

From here on I will try my best to thank the people that were close to me (not only at work) during these wonderful three years in Denmark:

The first person I have to thank is my old friend and never colleague Tejal, without whom I would have never found the PhD project I was part of, and of course thanks for spending very nice time together.

Thanks to my supervisor, Mette Voldby Larsen. If I do not remember any stressful time during my PhD is mainly thanks to her great supervision. All the suggestions she gave at the start of my PhD about the organisation and scheduling of all the tasks involved in a PhD student's life proved to be perfect. I had learnt a lot from her and I am happy to be her first PhD student. Apart from being a great supervisor, she is one of the kindest people I have ever met.

Thanks to my co-supervisor Ole Lund, is not only a great scientist always ready to give very good suggestions when asked, but also a great group leader. I am not sure how he does, but if the immunological group at CBS is so nice and cheerful it is also thanks to him choosing the right people and leading the group. And of course a big thank you here goes to all members of the immunological bioinformatics group.

A big thank to John Damm Sørensen, who helped me a lot and without whom it would have been impossible to do my research, teaching me how to use CBS's supercomputer facilities. He was always ready for a nice chat after tiresome working days, and he also was one of the first people from whom I have learnt about Danish culture and history.

Thanks to Kristoffer for helping me whenever I needed, giving me a wonderful office and desk and introducing me to the "let's talk about it over a cup of coffee" philosophy.

Thanks to all the people working at the administration. Sometimes we do not realise it, but if we would have to deal with all the bureaucracy they deal with for us, we simply would not have time to do any research. A special thank to Lone, Karina and Dorthe for always being kind and helpful.

Thanks to Cecile and all the office helpers before her. Drinking a coffee (I did not do that much at CBS) or steaming your milk making a lot of noise is something that becomes normal at some point, but if the machine was not working, my day would have had a bad start. So, thanks to the office helpers for taking care of the small things that make CBS a comfortable place to work.

Thanks to Peter and Edita, my first office mates. Working in Peter's office I understood how it is to work in what is, in my opinion, the most busy office at CBS.

Thanks to Greg, Txema and Li for being great office mates. Thanks to Txema for not getting too mad at me for systematically leaving something on his desk. By the way, you should try to smile, some people at CBS think you are very grumpy, and I know you are not...that grumpy. Thanks to Li for just being always so kind and for the nice time in and outside CBS.

Thanks to Thomas, Sonny, Marcin, Marcelo, Ali, Oksana, Dhany, Edita, Francesco, Federico, Tammi.

Thanks to Professors Iida and Nakamura for hosting me in Osaka University during my external stay.

Thanks to Yukako, Junko and Nuccio for always making me feel home every time I go to Japan, and of course for being great people and friends.

Thanks to Daniele and Stefania for being my friends for so long and for all the nice time spent during my PhD, before and I am sure after. I am sure Denmark would have been different without you "sbranjisati".

Thanks to my neighbour and friend Karina, for making my life in Søborg less ordinary and for helping me in different situations, also some tragicomic ones.

Thanks to Yukako and Miki for helping me in designing the thesis cover.

And lastly, the biggest thank to my family and my nieces for always being close to me no matter how far I am living from them.

Papers included in the thesis

- **Salvatore Cosentino**, Mette V. Larsen, Frank M. Aarestrup, Ole Lund. *PathogenFinder - Distinguishing Friend from Foe Using Bacterial Whole Genome Sequence Data.* PLoS ONE 2013, 8:e77302.

- Mette V. Larsen, **Salvatore Cosentino**, Oksana Lukjancenko, Dhany Saputra, Simon Rasmussen, Henrik Hasman, Thomas Sicheritz Pontén, Frank M. Aarestrup, David W. Ussery, Ole Lund. *Benchmarking of Methods for Genomic Taxonomy.* [Under review in Journal of Clinical Microbiology]

- Mette V. Larsen, **Salvatore Cosentino**, Simon Rasmussen, Carsten Friis, Henrik Hasman, Rasmus L. Marvig, Lars Jelsbak, Thomas Sicheritz Pontén, David W. Ussery, Frank M. Aarestrup, and Ole Lund. *Multilocus Sequence Typing of Total-Genome-Sequenced Bacteria.* Journal of Clinical Microbiology 2012, 50:1355–1361.

- Ea Zankari, Henrik Hasman, **Salvatore Cosentino**, Martin Vestergaard, Simon Rasmussen, Ole Lund, Frank M. Aarestrup and Mette Voldby Larsen. *Identification of acquired antimicrobial resistance genes.* Journal of Antimicrobial Chemotherapy 2012, 67:2640–2644.

Abbreviations

AMR:	antimicrobial resistance
bp:	base pairs
CCD:	charge-coupled device
CDC:	Centers for Disease control and Preventions
CGE:	Center for Genomic Epidemiology
DBG:	De Brujin graph
DNA:	deoxyribonucleic acid
dNTP:	deoxyribonucleic triphosphate
EBI:	European Bioinformatics Institute
EMBL:	European Molecular Biology Laboratory
FSG:	First Generation Sequencing
HGT:	Horizontal Gene Transfer
INSDC:	International Nucleotide Sequence Database Collaboration
IT:	information technology
MLST:	Multi-locus Sequence Typing
MRSA:	methicillin-resistant *Staphylococcus aureus*
NGS:	Next-Generation Sequencing
NIH:	National Institute for Health
OLC:	overlap layout consensus
PCR:	polymerase chain reaction
PacBio:	Pacific Bioscience
SNA:	single-nucleotide addition
SMRT:	single-molecule real time
USB:	universal serial bus
WGS:	Whole Genome Sequence

1 INTRODUCTION

Every year more than 7 millions deaths are the direct cause of bacterial infectious diseases. An estimated 0.2 million people die of pertussis and 1.3 millions of tuberculosis, while diarrhea, causing more than 2.5 million casualties, is one of the leading causes of death by infection worldwide [5].

To treat patients with bacterial infections that could not be cured using drugs selected based on the symptoms, is important to grow and isolate the bacteria responsible for the infection, identify its species, determine its pathogenic potential and assess its resistance to antimicrobial drugs.

Characterisation of bacteria are currently performed by clinical microbiologists using species-specific methodologies developed over decades, which are complicated and expensive in terms of both time and money. The characterization of the bacterium under investigation can take from days (for fast growing species like *Escherichia coli*) to months for slow-growing bacteria like *Mycobacterium tuberculosis*. Time is very important in case of infections and it can make the difference in saving the patient's life.

Apart from causing sickness and - in the extreme cases - the death of the infected person or animal, one of the main characteristics of infectious diseases is their ability to spread and cause bacterial infection outbreaks.

A disease outbreak is the occurrence of a given disease in excess of what would normally be expected in a determined community or geographical area. An outbreak can arise in a restricted geographical area and potentially extend to different cities or even countries causing a global epidemic. Outbreaks can last from days to months [38] and even years. Bacterial outbreaks usually involve novel bacterial strains that have acquired new genes or combinations of pathogenic features that make the diseases they are causing very difficult to treat, and aid the spread of the bacteria.

With the rapid advancement of genome technologies and with the consequent decreasing price of bacterial whole genome sequencing, new ways of characterizing bacteria and performing epidemiological surveillance has arisen. Modern high-throughput DNA sequencers can provide, with relatively low prices and reduced time, partially complete DNA sequences of disease-causing microbes that can be used by researchers to get insight on the mechanisms making the microbe under investigation pathogenic.

With additional improvements to high-throughput sequencing technologies and further simplification in sample preparation, DNA sequencing and related bioinformatics tools are likely to replace conventional culture-based and molecular typing methods to provide quicker and more reliable clinical diagnosis and data to be used for monitoring the epidemiology of infectious diseases [21].

Provided there is information sharing by all clinical and public health laboratories, these genomic tools could give life to a global system of linked

databases of bacterial genomes that would ensure more efficient detection, prevention, and control of local, emerging, and other infectious disease outbreaks worldwide.

The 4 manuscripts accompanying this thesis and composing its main 3 chapters will focus on the following 3 tasks using bacterial WGS: identifying the species of a clinical isolate; testing its properties, such as resistance to antibiotics, virulence, and potentially novel features involved in pathogenicity; characterization of bacterial population though bioinformatics typing methods and their application in epidemiology surveillance.

1.1 From Watson and Crick's double helix to usb-stick sized sequencers

The year 2013 marks the 60th anniversary of the discovery of the DNA double helix structure [72] by the biologists Francis Creek and James Watson in 1953 (Figure 1.1). From this discovery, which started the modern era of biology, 12 years passed before the first nucleic acid molecule was sequenced from *Escherichia coli* in 1965 [31], and 15 before the first experimental finding of a DNA sequence by Wu and Kaiser [75, 37]: It took 5 years for Wu and Kaiser to complete a 12 bases long DNA sequence from bacteriophage lambda [75], which also represents the first example of DNA sequencing in the history of science.

The technique that was crucial in the history of modern sequencing technologies was introduced in 1975 by Sanger under the name of the 'plus and minus method' [58, 61] . Sanger soon refined his technique by the introduction of the 'dideoxy method' [62], which is the foundation of the so called Sanger sequencing or First Generation Sequencing (FGS), and also the technique that has dominated the DNA sequencing scene during the last 30 years. The first nucleotide sequence obtained using the Sanger method was the genome of the *phiX174* bacteriophage [59, 60]. Later, right after the genomes of *Mycoplasma genitalium* and *Haemophilus influenzae* were completed in 1995 [26, 27], scientists started considering the possibility of studying the pathogenesis of bacteria based on their genome sequences [28].

The most important milestone achieved using FGS was the sequencing of the human genome [39, 70]. This was a 13-years long project, which was completed in 2003 [18] and which gave start to the era of genomics that we are still living in at present. In this introduction we will not describe the Sanger sequencing method, and interested readers are directed to related articles [47, 33].

In this short overview on sequencing technologies we will describe how the main instruments available in market work and spend a few lines on the future development in high-throughput sequencing technologies.

Second-Generation or Next-Generation Sequencing (NGS) includes various technologies that rely on a combination of template preparation, sequencing and imaging, together with genome alignment and assembly meth-

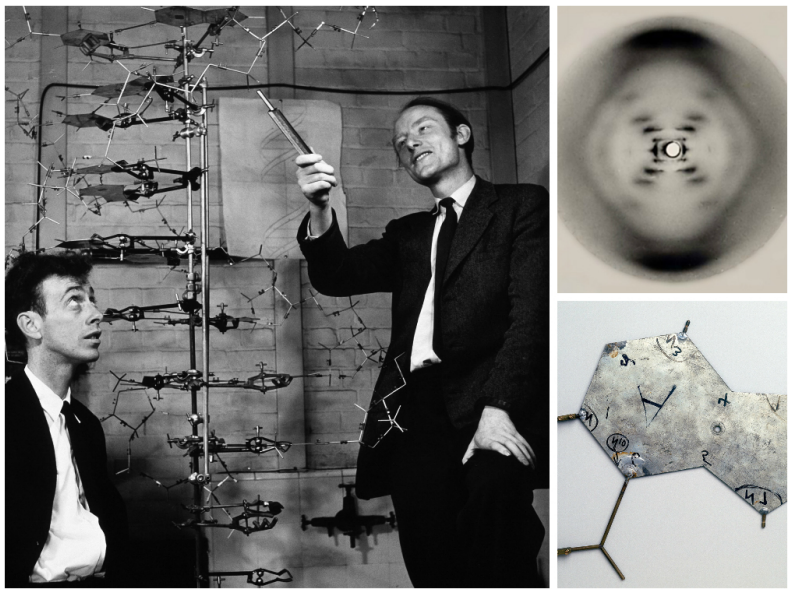


Figure 1.1: (left) Watson and Crick and their model of the DNA double helix (1953). (top-right) X-ray diffraction of the DNA taken by Raymond Gosling in the lab of Rosalind Franklin in 1952. (bottom-right) Aluminium template, representing the base adenine (A), part of Crick and Watson's model of DNA.

ods. NGS has changed the way of doing applied and clinical research, and its potential applications are only limited by the imagination of the users.

The primary advantage offered by NGS is the ability to produce gigantic amounts of short-reads per sequencer run, with constantly decreasing costs. One of the side effects of this sequencing revolution is the constant increase of the data to be stored, with the European Bioinformatics Institute (EBI) at present storing 2 petabytes (10^{15} bytes or 1000 Terabytes) of genomic data. This amount is more than doubling every year since the advent of NGS in 2008 (Figure 1.2) [46]. The volume of data produced by biologists is reaching quantities that were once the domain of high-energy physics and astronomy researchers. The second side effect, caused also by the short length of the reads combined with the huge amount of data produced per instrument run, is the increasing need for powerful computational infrastructures to process and analyze the generated experimental data.

NGS technologies are being applied in many fields of both basic and clinical research, for example in gene-expression studies, where these new technologies are slowly taking the place of microarrays, identifying transcripts without prior knowledge of a specific gene and providing information about sequence variation in identified genes [73]. Re-sequencing of human genomes is probably one of the most common applications of NGS, with projects like the "1000 Genomes Project" [7], started in 2008 and completed in 2012

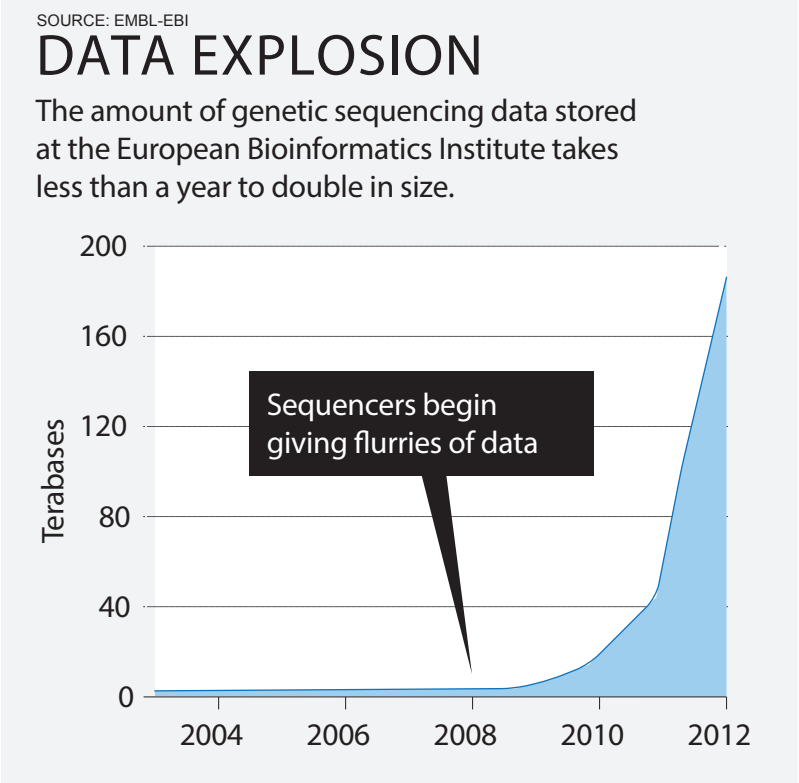


Figure 1.2: increase of the quantity of genomic data stored in EMBL-EBI servers before and after the NGS sequencers started being affordable for most researchers.

[10], in which a team of researchers from different countries sequenced the genomes of more than 1000 participants from different ethnic groups, creating a huge sequence dataset and a refined human genome map freely available to the scientific community and public. In this work we will focus on the use of high-throughput sequencing technologies for bacterial whole genome sequencing.

Sequencing platforms available in the market can be separated in **template amplification platforms** and **single-molecule sequencers.** Template amplification platforms depend on the creation of clonally amplified templates, and all of the NGS sequencers, including Roche 454, Illumina and Ion Torrent instruments belong to this class of sequencers. The general workflow for these sequencers involves the following 3 steps: 1) library (or template) preparation; 2) template amplification; 3) sequencing.

Library preparation starts with the extraction and purification of genomic DNA. For single-end shotgun sequencing [64, 8], a fragmentation step is needed, and depending on the application and platform, the length of the fragments can be from 150 to 800 base pairs (bp).

Template amplification platforms also support **mate-pair** and **paired-end** sequencing. In mate-pair sequencing DNA fragments of given lengths (e.g. 3 kbp, 8 kbp or 20 kbp) are joined together in circular molecules and fragmented again before adaptors are added to the fragments flanking the joins. In paired-end sequencing fragments are sequenced from both the 3' and 5' ends. From mate or paired-end reads is possible to obtain higher quality genome assembly than those attainable from single-end reads.

Amplification. In NGS platforms amplification is performed by immobilizing millions of spatially separated template fragments on to a solid surface, which can be a flow cell (Illumina platform), ion sphere particles (Ion Torrent) or solid beads (454 and SOLiD). In SOLiD, 454 and Ion Torrent platforms it is performed using emultion PCR [22], while in Illumina platforms the amplification is performed using bridge amplification (Figure 1.3).

NGS platforms use different chemistries to do the sequencing and different approaches (usually imaging) to read the sequences.

Roche 454 platforms use single-nucleotide addition (SNA). In SNA, at each cycle, one of the four possible dNTP (deoxynucleotide triphosphate) is, in turn, flown across the template DNA fragment. If there is complementarity between the added dNTP and the nucleotide in the next available position in the template, the base is attached to the extending DNA strand and hydrogen ions and pyrophosphate are released. This technique is known as pyrosequencing [56, 44] because the pyrophosphate released when a nucleotide is added to the extending strand is converted into visible light, which is measured using high resolution charge-coupled device (CCD) camera sensors. The output is a binary file in the standard flowgram format (SFF), containing the measured light peaks, that can be converted to a FASTQ [17] file containing the raw-reads.

Ion Torrent platforms also use the SNA method but, differently from the other NGS sequencers, they do not rely on imaging for reading the sequences and use instead a silicon chip to spot the positively charged hydrogen ions released every time a dNTP is added to the growing DNA strand [57].

Illumina platforms use an approach called cycling reversible termination developed by Solexa [12]. In each sequencing cycle, fluorescently labeled nucleotides are sprinkled across the flow cell. Nucleotides that are complementary to the next available position on the template strands, are incorporated in the growing DNA strands, and a laser is used to determine the incorporated nucleotide based on the fluorescent moiety. A new cycle starts after washing of the flow cell.

SOLiD platforms use the sequencing by ligation [69], in which fluorescent probes are iteratively hybridized and ligated to complementary nucleotides in the template strand at the 5' end of the growing strand, before fluorescence imaging is used to detect the ligated probes.

Single-molecule sequencers have a more straightforward preparation of templates than NGS instruments, requiring less DNA and without the need of PCR. A single-molecule sequencer that available at present in the market

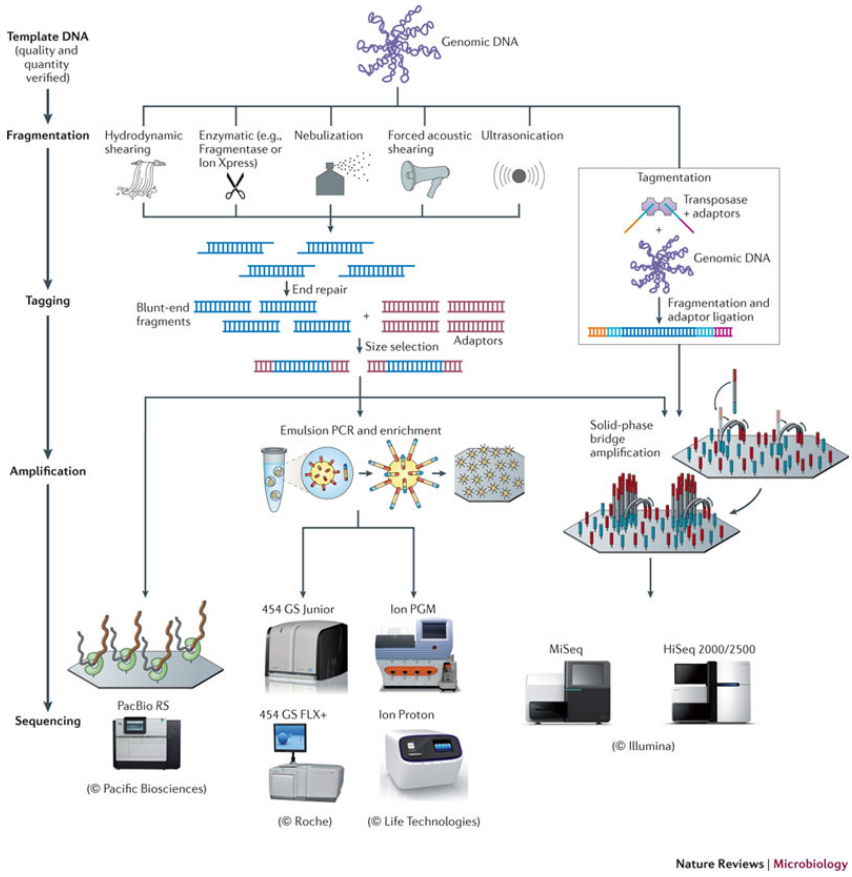


Figure 1.3: steps needed to run a sequencing experiment using the main high-throughput sequencing platforms available in the market. The schematic associates, to each platform, the sample preparation and template amplification procedures needed before sequencing. Illustration from [43], ©2012 Macmillan Publishers Limited

Figure 1.4: the MinION sequencer, announced by Oxford Nanopore Technologies, is the smalled instrument in the history of sequencing and can be used by simply plugging it to the USB port of a computer. ©2008-2013 Oxford Nanopore Technologies.

is the PacBio RS from Pacific Bioscience. The sequencing approach used by PacBio models is known as single molecule real time sequencing (SMRT) [24]. One of the striking characteristics of SMRT sequencers is the high length of the generated reads, with an average of 3000 bases while one of its limitations is in its rather high error rate, which it has recently been shown to be possible to reduce by post correcting the sequenced data [36].

Figure 1.3 shows the main high-throughput sequencing platforms available in the market today, and associates, to each platform, the sample preparation and template amplification procedures required before sequencing.

In February 2012, the sequencing community was shook by the demonstration, from a UK startup company called Oxford Nanopore Technologies, of a new revolutionary sequencing platform. Oxford Nanopore presented their two nanopore-based sequencers (the GridION and MinION), capable of delivering extremely long reads (they said there is not theoretical limitation on the maximum length) at very cheap costs (less than 1000$ for a human genome) and directly from blood samples. MinION (Figure 1.4) was the most admired also because of its USB memory-stick size [25]. The UK-based company recently started selling the sequencers to researches who would like to test the technology.

Nanopore DNA sequencing method involves passing single DNA strands through tiny (nano) pores. The pores (through which current is flowing), with a diameter of about 2 nanometers, will identify the bases by measuring the change in current flow caused by the bases passing through it. Recently

US National Institute for Health (NIH) has granted about $17 million to 8 research teams to conduct studies on the use of nanopore technology for DNA sequencing [2] which is another hint of the possibility for this revolutionary technologies to become a reality in the near future.

Together with the limitation related to the sequencing costs, nanopore sequencing, producing very long reads, could solve many of the problems related to de-novo assembly by directly generating almost complete genomes in a single instrument run. The solution to this problem will reduce also the costs related to expensive computer equipments required at present to assemble genomes from the output of sequencers currently in the market.

1.2 WGS data analysis and related applications

The fact that many sequencing instruments are available and competing in the market is an advantage for the final users, since the companies have to improve their technologies and reduce the prices if they want to keep their market shares. One of the downside is the fact that users will have to deal with different data formats, and giving the pace at which sequencing technologies are evolving, users are struggling to understand what are the best tools or pipelines to analyse their data, creating a scenario in which it is relatively easy and cheap to obtain WGS data, but everyday is getting increasingly complicated to analyse and interpret the data. High-end sequencers produce a quantity of data by each run for which storage and analysis would be too demanding for the average microbiology lab, whereas bench-top sequencers deliver quantities of data per run which better suits the IT equipment of the average microbiology research lab.

Once the run of the sequencer is completed, a preprocessing of the short reads, involving a quality filter and sequencing error correction, is performed before the assembly starts. Preprocessing aims at the removal of low-quality and erroneous reads from the data, with the final purpose of improving the quality of the subsequent assembly. Different technologies have different types of errors mainly derived by the sequencing chemistry they use, and correcting these errors is very important since de-novo assembly is particularly sensitive to these errors [53]. After the preprocessing of the data, algorithms are used to join the reads in a logical fashion to build the final sequences (contigs) that will compose the bacterial genome, in a process commonly called reads assembly. Depending on the kind of study a *reference assembly* or a *de-novo assembly* of the WGS data is performed.

Reference assembly basically consists in mapping the reads (through alignment algorithms) to the reference genome in order to identify genetic differences in the compared (usually highly related) genomes. Reference genomes are used, for example, to complete genomes of bacteria that have been partially assembled (draft genomes), or to study how different a given strain is in comparison to other strains of the same species. The latter is very important for epidemiological studies.

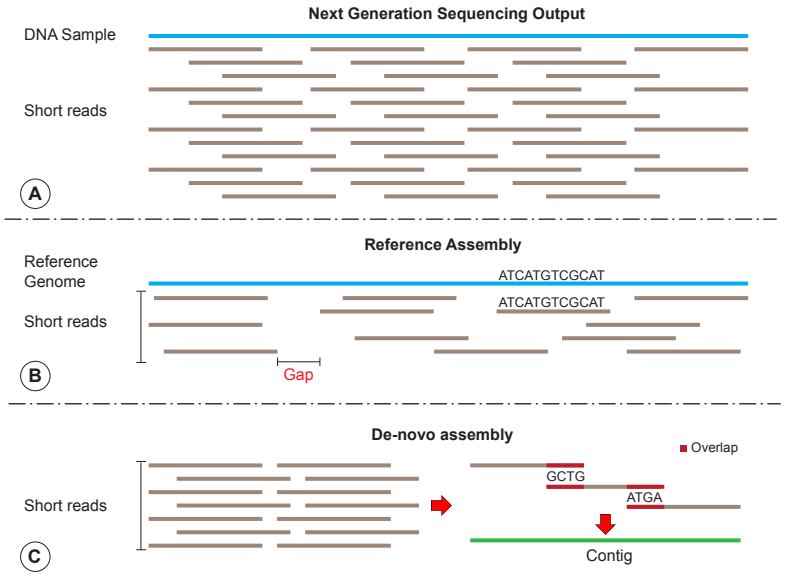


Figure 1.5: Next generation sequencing output and genome assembly approaches. A) Shows a simplified example of the reads produced by the run of a NGS sequencer. B) Depicts how reference assembly works. The short reads are aligned against the reference genome (cyan long segment) and the parts of the reference to which no short read is mapping will generate gaps in the final assembly. C) Shows a simplified example of de-novo assembly, where a group of short reads (on the left) are sorted and aligned to each other to generate a contig sequence.

In a scenario in which the bacterium under study is unknown, we may want to perform a de-novo assembly of the WGS data, which consists in the reconstruction of the genome sequences without the help of any other information apart from the short reads. In de-novo assembly, reads alignment is done by comparing the reads with each other, or by overlapping k-mers (reads sub-sequences of k length). The final purpose of these alignments is the creation of contiguous sequences called contigs. Figure 1.5 depicts simple examples of both reference and de-novo assemblies. In an ideal situation the assembly will produce a single sequence containing, for example, the DNA composing the chromosome of the sequenced strain. Unluckily this is not so easy, in fact while short reads are good for reference assembly they are very difficult to assemble de-novo. Most de-novo assemblers are based on graph theory and represent overlaps (of either reads or k-mers) through vertices and edges in the graph that, after its completion, will be used to create the contigs. De-novo assemblers can be divided in Greedy, Overlap Layout Consensus (OLC) and De Brujin Graph (DBG) type, depending on the approach they use to preform the assembly [51].

A typical difficulty for de-novo assembly algorithms is in finding a path in the overlap graph that passes through each vertex only one time (Hamil-

tonian path) or through each edge only once (Eulerian path). This problem is the reason why graph based assembly algorithms are very sensitive to sequencing errors [53]. Another limitation of these algorithms is related to the high amount of computer memory, and the powerful processors they need to be executed. This problem is both due to the nature of the mathematical problem these algorithms are trying to solve and to the high amount of short reads generated by NGS sequencers.

The OLC method is composed by the following 3 phases: 1) an overlap phase [52], in which all reads are aligned with each other to find overlaps, taking into account the minimum k-mer and overlap size, which will affect the quality of the final contigs; 2) a layout phase [51], in which a graph is built and iteratively updated, corrected and optimised; 3) a final consensus phase in which the final assembly is produced by means of multiple alignments. Among the assembly tools based on the OLC approach there are MIRA [16], ARACHNE [11], Celera Assembler (CABOG) [50], the PacBio-specific Allora and the commercial Newbler from Roche.

The DBG method [53] consists in building graphs in which each vertex represents a k-mer that will appear in the graph only one time. Reads are decomposed into k-mers that are then mapped to the graph, reducing the mathematical problem to the calculation of a Eulerian path [54] which can be solved in linear-time [67] even with millions of reads. Among the disadvantages of the DBG method there is the high amount of memory needed to execute the algorithm and the inability to calculate a priori the k-mer length to be used, which requires the algorithm to be run many times with different k-mer sizes in order to find the optimal settings. Most of the available genome assembly tools based on DBG are used to assembled reads from Illumina and SOLiD sequencers, and among these are ALLPATHS-LG [14, 29], SOAPdenovo [42], Velvet [77, 78], which is one of the most widely used, and Ray [13], which can assemble genomes by mixing 454 and illumina reads.

The Greedy approach iteratively alignins reads to identify the fragments with the largest overlap and merge them. This approach is used to assemble Sanger sequencing data, and among the tools based on it there are TIGR Assembler and PHRAP [55]. Some tools (SHASSAKE [71] and VCAKE [35]) use this approach to assemble short reads, but given their really high computational needs they never become popular among users.

Improving the quality and performances of genome assembly algorithms is one of the main challenges in computational biology. Given the importance of the topic and the high variety of available tools and pipelines communities, for testing, improving and evaluating genome assembly tools arose [23].

Even though genome assembly is one of the most important steps in WGS data analysis it is just the starting point for the characterisation of the bacteria under investigation. In fact after the genome assembly many other algorithms and tools will need to be applied for annotation, antibiotic resistant and pathogenic genes identification, typing, phylogenetic comparison and all other tests needed for the characterisation of the sequenced strain.

1.3 What is a Human Pathogen?

A pathogen is usually defined as a microorganism that causes, or can cause, a disease in a host. However, this definition is insufficient for describing what a pathogenic bacteria is, which takes us to an ongoing debate that dates back to 1880, when the German microbiologist Robert Koch defined a postulate to assess the pathogenicity of bacteria in human. Back then it was considered enough for a bacterium to have some virulence factors, for instance toxins, that could make a person sick, for it to be identified as human pathogenic. Koch's postulate started becoming unreliable for the assessment of bacterial pathogenicity during the middle of last century when the first antimicrobial drugs were introduced and bacteria that were considered harmless at the time, like *Staphylococcus aureus* and *Clostridium difficile*, started causing serious infections.

The species just mentioned are usually classified today as *opportunistic pathogens*. The use of broad spectrum antibiotics is a major cause of opportunistic infections altering the population of commensal and pathogenic bacteria alike, while at the same time facilitating the proliferation of microbes resistant to the administrated drugs [20], which will grow uncontrolled and cause a new infection. Opportunistic infections also occur in people with compromised immune system due, for example, to hereditary genetic alterations. An example is the cystic fibrosis, in which the opportunistic pathogen *Pseudomonas aeruginosa*, normally found in healthy people, is the bacteria causing the infection.

Acquired antimicrobial resistance (AMR) is nowadays responsible for hundreds of thousands of infections worldwide [6]. Accordingly, it is considered one of the top health challenges facing the 21^{th} century. AMR causes by multiple reasons among which: the misuse of antibiotics for treating human diseases [48, 1]; the use of high doses of antibiotics in livestock's food [45]; the use of antibiotics in healthcare facilities [3]. Figure 1.6 shows the main reasons behind the development of antibiotic resistance and how it can spread to humans.

Another class of bacteria that are worth to mention are the *zoonotic pathogens*. These pathogens usually infect animals, but the infection can be transmitted to humans being in direct contact with the pathogen's hosts, causing a zoonotic infection (or zoonosis). Zoonotic infections can also be caused by the consumption of contaminated food or water. In a systematic study it has been shown that out of more than 1400 bacteria infecting human, 66% were zoonotic [66]. Among the many zoonotic infections there are those caused by the methicillin-resistant *Staphylococcus aureus* (MRSA). The bacteria is usually found in livestock (cattle, pigs and poultry), from which can be transmitted to human through direct contact or through the consumption of contaminated meat, causing the zoonosis. Apart from being zoonotic MRSA is also one of the main causes of nosocomial infections, which are very difficult to treat also due to its antibiotic resistance. Recently MRSA infections are becoming a big threat in Denmark [1] and worldwide. The most

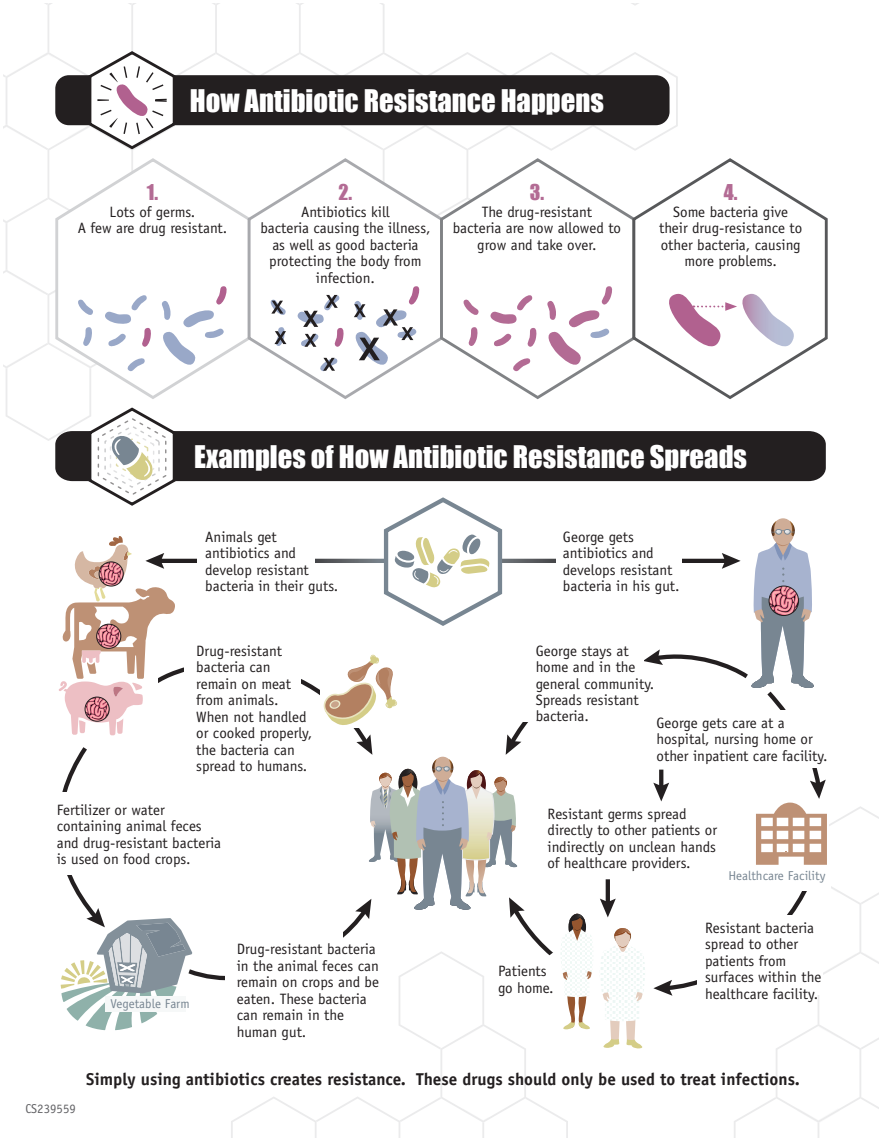


Figure 1.6: Infographic showing how antibiotic resistance happens and how multi-drug resistant microbes can spread to humans. ©Centers for Disease Control and Prevention (CDC).

Figure 1.7: (left) *Trionfo della Morte* (*Death's Triumph*), a 1446 fresco from Sicily (Italy). Historians say that bubonic pest epidemy started in Mongolia and reached Europe through trading boats from the Black Sea, that brought, together with their goods, infected people and rats to the harbors of Sicily and Genova (Italy); (right) the oriental rat flea (*Xenopsylla cheopis*) was the main vector though which the epidemy spread. This parasite fed on rats infected with the pathogenic bacteria *Yersinia pestis* (causing the bubonic pest), and trasmitted the microbe to humans and other mammals when feeding on them.

famous and deadly zoonosis was the *Black Death*, a bubonic plague outbreak, causes by *Yersinia pestis*, that killed more than 200 million people in Europe during the 14^{th} century (Figure 1.7), and is still considered today the most devastating pandemic in human history.

From above, it is already clear how complicated it can be to classify a given bacterial strain as either pathogenic or commensal (Figure 1.8).

Although it is challenging is to predict the emergence of new pathogenic strains or simply separate pathogenic from opportunistic and commensal strains, there are nevertheless examples of models for the prediction or estimation of pathogenicity factors through the analysis of DNA sequences [9, 34, 19]. Apart from the difficulties deriving from the definition of pathogenic bacteria itself, one of the main difficulties in the creation of prediction models is related to the lack of well curated databases for microbes known to be human pathogenic. Luckily, projects like the 100k pathogen genome project [4], in which researchers are trying to sequence 100 thousand genomes of bacteria and viruses causing foodborne infections, will help in having better collections of strains that could be used for building reliable prediction models.

1.4 Web-based Tools for the Characterisation of Prokaryotes

With the constantly decreasing cost of sequencing and the continuous development of new tools for high-throughput sequencing data analysis, many

Figure 1.8: A) advertise for a probiotic dairy product made by fermenting a mixture of skimmed milk with a special strain of the bacterium *Lactobacillus kasei* of the *Shirota* strain; B) a puppet representing the multi-drug resistant *Staphylococcus aureus* (MRSA) (the cloak symbolizes its antibiotic resistance); C) a microscope image of an attenuated strain of the zoonotic *Mycobacterium bovis* (closely related to *Mycobacterium tuberculosis*, which causes tuberculosis in human), which causes tuberculosis in cattle and can potentially be transmitted to human causing tuberculosis.

microbiology labs and healthcare institutions are gradually starting using DNA sequencing for the characterisation of the bacterial strains they are studying, and within a few years these labs are expected to use sequencing on a daily bases. Soon the limiting factor in genomic research will be not to obtain the sequenced data but to analyse it, which directly relates to the problem of training the lab's staff [15].

The Center for Genomic Epidemiology (CGE) aims at the creation of web-based tools for the rapid analysis of sequence data, to be available for free to the global scientific and medical community. The web-tools target users that are scientists and clinicians with limited computer skills, whom would find it very difficult to use not user-friendly bioinformatics software.

Less than two years from the publication of the first CGE service (MLST) [40], the answer from scientific community has been more than positive, with the number of users and served jobs more than tripled in just one year (Figure 8) , with users from more than 60 countries worldwide (Figure 9). At present (November 2013), the four published CGE web-tools [40, 76, 41, 19] are serving more than 1500 jobs per month, with MLST and ResFinder [76] being the most used tools. The users are mainly from developed countries

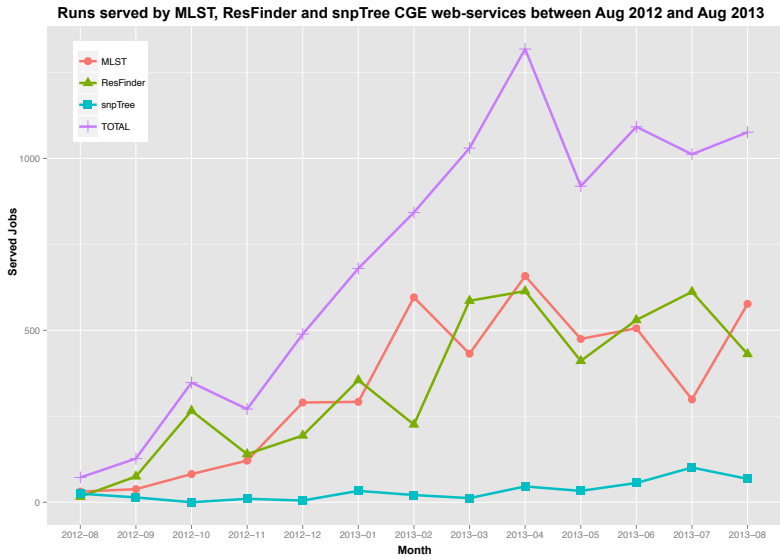


Figure 1.9: jobs served by MLST [40], ResFinder [76] and snpTree [41] web-services from CGE in the period between August 2012 and August 2013.

(30% US, 15% UK), but 10% of the users are from developing countries like Philippines, Thailand, Indonesia and other countries in Africa and South America, for which using our tools for free can save important money for their research, considering the costs of the wet lab versions of both MLST and antimicrobial resistance assessment.

At present the CGE tools are able to de-novo assemble and analyse bacterial WGS data, and during the next 2 years a complete pipeline should allow users to have a fast and complete characterisation of the bacterial strains they are studying. The pipeline will be able to do taxonomy identification and typing on the input DNA sequences, as well as identification and prediction of phenotypic traits, e.g., antibiotic resistance and pathogenicity features (tools which are already available as single web-services). Together with the analysis needed to scientists and medics in order to understand what is the bacteria causing a given infection, the pipeline will implement tools [41] to be used for epidemiological studies, which would be very helpful to detect potential outbreaks.

Three [40, 76, 19] of the four articles included in this thesis result from the development of algorithms, and the related web-services, offered by CGE.

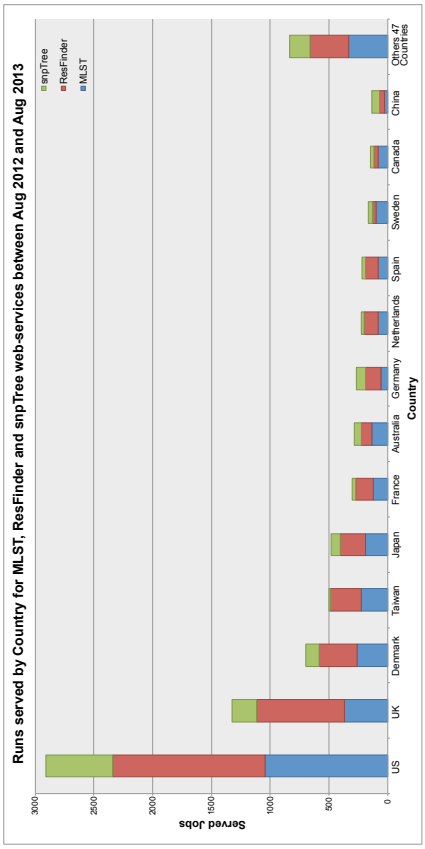


Figure 1.10: jobs served by country for MLST [40], ResFinder [76] and snpTree [41] web-services from CGE between August 2012 and August 2013. Users come from more than 60 countries, with developing countries like Philippines, Thailand, Indonesia and India accounting for about 10% of the total use of the CGE web-servers (Nov 2013).

2

BACTERIAL PATHOGENICITY PREDICTION USING WHOLE GENOME SEQUENCE DATA

Although some of the most studied bacterial strains can cause serious diseases that could sometimes kill the infected person, not all bacteria are dangerous and many of them are innocuous or even beneficial to human. With almost 2600 publicly available complete genomes and more than 1900 submitted to International Nucleotide Sequence Database Collaboration (INSDC) (*www.genomesonline.org*, November 2013), many researchers are using this invaluable information to study the mechanism behind bacterial pathogenesis.

Nevertheless, as discussed above, one of the main difficulties in making prediction models for the emergence of pathogenic bacteria are related to the definition of human pathogen bacteria itself. Among the difficulties in building these prediction models there is the high similarity between pathogenic and commensal strains for species like *Escherichia coli,* also due to Horizontal Gene Transfer (HGT) [30, 65]. Opportunistic pathogens, which are bacteria that can be found in healthy people without causing any infection but can be deadly for people with compromised immune system, are another challenging problem related to the creation of bacterial pathogenicity prediction models.

In the manuscript proposed in this chapter [19] we started from the work proposed in [9], and we modify and extend the method to work with all kinds of bacteria. Each bacterial strain was tagged as human pathogenic if at least one case of human infection, or infections of other mammals caused by a given bacterium, was found in scientific literature. The bacterium would be tagged as non-pathogenic otherwise. Strains infecting fishes or plants with no evidence of zoonosis were also tagged as non-pathogenic.

With the results obtained in this manuscript we shown that it is possible to make prediction models for bacterial pathogenicity provided the availability of a relatively high amount of complete genomes for each species and at the same time a complete dataset containing a wide variety of species. The implemented method is available as a free to use web-service (*http://cge.cbs.dtu.dk/services/PathogenFinder/*) part of the services offered by the CGE project.

Manuscript I

PathogenFinder - Distinguishing Friend from Foe
Using Bacterial Whole Genome Sequence Data

OPEN ACCESS Freely available online

PLOS ONE

PathogenFinder - Distinguishing Friend from Foe Using Bacterial Whole Genome Sequence Data

Salvatore Cosentino[1]*, Mette Voldby Larsen[1], Frank Møller Aarestrup[2], Ole Lund[1]

1 Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Kgs. Lyngby, Denmark, **2** National Food Institute, Technical University of Denmark, Kgs. Lyngby, Denmark

Abstract

Although the majority of bacteria are harmless or even beneficial to their host, others are highly virulent and can cause serious diseases, and even death. Due to the constantly decreasing cost of high-throughput sequencing there are now many completely sequenced genomes available from both human pathogenic and innocuous strains. The data can be used to identify gene families that correlate with pathogenicity and to develop tools to predict the pathogenicity of newly sequenced strains, investigations that previously were mainly done by means of more expensive and time consuming experimental approaches. We describe *PathogenFinder* (http://cge.cbs.dtu.dk/services/PathogenFinder/), a web-server for the prediction of bacterial pathogenicity by analysing the input proteome, genome, or raw reads provided by the user. The method relies on groups of proteins, created without regard to their annotated function or known involvement in pathogenicity. The method has been built to work with all taxonomic groups of bacteria and using the entire training-set, achieved an accuracy of 88.6% on an independent test-set, by correctly classifying 398 out of 449 completely sequenced bacteria. The approach here proposed is not biased on sets of genes known to be associated with pathogenicity, thus the approach could aid the discovery of novel pathogenicity factors. Furthermore the pathogenicity prediction web-server could be used to isolate the potential pathogenic features of both known and unknown strains.

Citation: Cosentino S, Voldby Larsen M, Møller Aarestrup F, Lund O (2013) PathogenFinder - Distinguishing Friend from Foe Using Bacterial Whole Genome Sequence Data. PLoS ONE 8(10): e77302. doi:10.1371/journal.pone.0077302

Editor: Jason D. Barbour, University of Hawaii Manoa, United States of America

Received August 2, 2013; **Accepted** September 9, 2013; **Published** October 28, 2013

Funding: This work was supported by the Center for Genomic Epidemiology (http://www.genomicepidemiology.org) at the Technical University of Denmark and was funded by grant 09-067103/DSF from the Danish Council for Strategic Research. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: salvocos@cbs.dtu.dk

Introduction

Every year more than 15 millions deaths are the direct cause of infectious diseases, many of which are due to bacterial infections. Each year an estimated 1.3 million people die of tuberculosis and 0.2 millions of pertussis, while diarrhoea accounts for more than 2.5 millions deaths, and is one of the leading causes of death in worldwide [1]. But not all bacteria are dangerous and many of them are innocuous or even beneficial to human. The gut of a healthy adult human contains thousand of different microbial species, many of which are beneficial to their host, providing functions for nutrition and development, and regulating the immune response [2,3]. Nevertheless some bacterial species, like *Escherichia coli*, also include extremely deadly strains, causing for example diarrhoea, urinary tract infections, septicaemia etc. Thus identifying pathogenic strains and understanding the biological mechanisms that cause them to become pathogenic is highly important to perform timely interventions and design control strategies, including interventions such as restrictions on contaminated food products, isolation of patients, correct treatment and development of targeted vaccines.

Ever since the 1880s the pathogenicity of bacteria have been assessed using Koch's postulates, for human pathogens using animal models. During the last 2 decades many discoveries have shown that Koch's postulates are not enough to decide if a given bacteria is pathogenic or not. The existence of diseases caused by bacteria that cannot grow in pure culture medium [4,5], the discovery of polymicrobial diseases [6], the role of metagenomic microbiota in chronic diseases [7], and last but not least, the discovery of Horizontal Gene Transfer (HGT) responsible for the swapping of genetic material between bacteria [8] (regardless the pathogenicity), are all cases in which the postulates have shortcomings. Already during his work with *Vibrio cholerae* Robert Koch himself discovered the shortcomings of animal models for correctly identifying human-specific pathogens. Thus, the use of animal models is not always reliable in defining if a given bacteria is human pathogenic. Moreover, assessing the pathogenicity by means of animal models or epidemiological studies is both time-consuming and expensive.

Among the molecular features that a bacterium needs to infect and survive inside its host [9] are exotoxins, endotoxins, two components systems [10], adherence factors, secretion systems (I to IV type) [11], through which bacteria can inject their toxins into its hosts cells [12]. Plasmids, secretion systems, and antibiotic resistance genes are commonly present in both commensal and pathogenic strains, while toxins are usually only present in pathogenic strains. There are many databases available containing genes encoding toxins and virulence factors along with other genes traditionally associated with pathogenicity [13,14].

One of the ways to classify a bacterium as human pathogenic using bioinformatics was (and still sometimes is) to look for some of these features in the genome of the isolate under investigation.

Table 1. Training, test data and model parameters.

	Training Set			Test Set			Model Parameters			
Model Name	**Pathogenic**	**Non-pathogenic**	**Total**	**Pathogenic**	**Non-pathogenic**	**Total**	**MinORG**	**LT**	**HT**	**Zthr**
TM-Alphaproteobacteria	29	60	89	11	28	39	2	0.15	0.6	10.43
TM-Betaproteobacteria	26	26	52	10	22	32	2	0.3	0.9	0.55
TM-Epsilonproteobacteria	17	5	22	16	2	18	2	0.4	1.0	−9.31
TM-Gammaproteobacteria	122	97	219	33	50	83	2	0.2	0.85	25.37
TM-Actinobacteria	27	44	71	24	36	60	2	0.0	1.0	−3.22
TM-Bacteroidetes	7	12	19	5	24	29	2	0.35	0.6	1.68
TM-Firmicutes	98	87	185	34	83	117	3	0.0	1.0	−2.85
TM-Tenericutes	6	8	14	5	9	14	2	0.0	1.0	−1.59
COMPL	40	174	214	17	40	57	2	0.0	1.0	−1.78
WDM	372	513	885	155	294	449	2	0.0	1.0	3.0

Training, test data and model parameters. The last 3 columns show the MinORG, LT and HT parameters used to create the pathogenicity families and build the model for each of the 10 models. *Zthr* is a threshold value, calculated for each model at the cross validation phase, which is used, given the final prediction score, to decide if the input organisms will be predicted as pathogenic or non-pathogenic. The parameters for each model are chosen after 5-fold cross-validation tests.
doi:10.1371/journal.pone.0077302.t001

Unluckily this approach is not always reliable, partly because of HGT, which causes these features to be exchanged among pathogenic and innocuous strains of the same [15][16] or different species, an exchange which has been proved by the high amount of these features found in genomic islands [17]. Aside from the features directly associated to pathogenicity, there are also virulence "lifestyle" genes, important for the bacteria to survive inside the host and evade its immune system response [18][19], and genes that are, for example, needed to activate other genes, which are important in the processes of pathogenesis, even though they do not directly determine virulence. All the issues related to the prediction of bacterial pathogenicity based on phylogeny has caused researchers to look for different solutions.

Table 2. MCC on cross validation and independent test-set.

Organism subset	**5-fold CV**	**TM or COMPL**	**WDM**
All Bacteria	0.847	0.736[3]	0.758
α-proteobacteria	0.949	0.886	0.873
β-proteobacteria	0.923	0.855	0.79
ϵ-proteobacteria	0.741	0.686	1.0
γ-proteobacteria	0.825	0.666	0.661
Actinobacteria	0.681	0.816	0.826
Bacteroidetes	0.889	0.535	0.383
Firmicutes	0.915	0.756	0.785
Tenericutes	0.866	−0.344	0.0
Remaining Organisms[1]	0.940	0.793	0.877[2]

Column 2, the MCC obtained in the 5-fold cross validation (CV) by each of the 10 models. Column 3, the MCC of the individual TM models and the COMPL model (last line) when tested on independent test data from the corresponding phyla/classis. Column 4, the MCC of the WDM model when tested on independent test data from specific phyla/classis.
[1]Organisms of phylum/class for which no TM model is available were tested using COMPL model. COMPL was trained on all organisms from classes or phyla for which only either pathogenic or non-pathogenic strains were available.
[2]MCC for WDM on the same test-set used for COMPL.
[3]Overall MCC for all the TM models and the COMPL model.
doi:10.1371/journal.pone.0077302.t002

The development of whole genome sequencing may open novel ways of predicting pathogenicity in bacterial species. In 1995 the genomes of *Mycoplasma genitalium* and Haemophilus influenzae [20,21] were completely sequenced, and scientists started considering the possibility of studying the pathogenesis of bacteria based on their genome sequences [22]. This was the start of a revolution that has been continuing during the last decade with the advent of Second-Generation or Next-Generation Sequencing (NGS), leading to a continuous decrease in sequencing costs and a fast development of sequencing technologies. At present, many different high-throughput sequencing systems are available [23–25] and the number of completely sequenced bacteria amount to almost 2,400 including more than 1,800 that have been submitted to the International Nucleotide Sequence Database Collaboration (INSDC) (www.genomesonline.org, May 2013).

A few methods have been proposed which make use of Support Vector Machines (SVM), BLAST or other bioinformatics tools to search for pathogenic features [26,27] or predict bacterial pathogenicity [28] by searching in pre-computed databases of genes associated with pathogens. One shared aspect among these methods is the fact that they restrict their search to well known pathogenic features, missing out on the information that may be contained in the many genes with unknown function. Furthermore, the methods ignore genes that could be shared and specific among non-pathogenic organisms. When bacteria become pathogenic through HGT their lifestyle change and some of the genes may be inactivated or even lost to adapt to the new lifestyle [29,30]. These genes are still present in non-pathogenic bacteria and hence could be used, together with the genes associated to pathogenicity, to separate dangerous bacteria from harmless ones.

As an alternative to the above mentioned prediction methods, we here developed a novel approach, building on a previous study [31]. In this study we selected groups of genes which are frequently found either in human pathogenic bacteria or in the innocuous ones, and show that this is more effective than using global similarity. Since we did not make any pre-assumption on the genes contained in our training-sets, we are able to identify new proteins associated to pathogenicity and also features shared among non-pathogenic bacteria. Moreover, our hypothesis-free approach gave us the chance to build, together with a phylogenetic-independent model using all the organisms we have, more specific models

Table 3. Top 10 ranking pathogenic protein families and annotated functions of their proteins for TM-Gammaproteobacteria model.

RANK	Z-score	P	N	Function
1	9.134	77	8	N-acetylmannosamine kinase (TCS)
2	8.500	49	0	Fimbrial proteins
3	8.170	62	6	Sialic Acid Transporter
4	8.158	53	3	Transposition helper protein
5	8.023	62	7	Acetyltransferase, type III secretion proteins
6	8.023	62	7	Macrolide-specific efflux, membrane protein
7	8.023	62	7	Type II secretion proteins
8	7.922	69	10	Unknown function, possible membrane proteins
9	7.906	60	7	Unknown function
10	7.855	53	4	Cythochrome b_{562}

P and N columns contain the number of pathogenic and non-pathogenic organisms in the protein family respectively.
doi:10.1371/journal.pone.0077302.t003

Table 4. Top 10 ranking non-pathogenic protein families and annotated functions of their proteins for TM-Gammaproteobacteria model.

RANK	Z-score	P	N	Function
1	−6.52	3	34	Protein-L-isoaspartate
2	−6.44	2	31	ThiJ/PfpI domain protein
3	−6.43	6	40	Anthranilate synthase component I
4	−5.98	6	36	8-amino-7-oxononanoate synthase
5	−5.92	5	34	Unknown function, putative transcriptional regulator
6	−5.82	0	21	Adenosylmethionine decarboxylase
7	−5.81	8	39	Unknown function
8	−5.80	2	26	Unknown function, probable condensation protein
9	−5.68	0	20	Nitrite transporter
10	−5.62	1	22	Glucose-galactose transporter

P and N columns contain the number of pathogenic and non-pathogenic organisms in the protein family respectively.
doi:10.1371/journal.pone.0077302.t004

grouping organisms at different taxonomic ranks to improve the predictions in species like *E. coli*, in which the high amount of shared genes among pathogenic and commensal strains makes it particularly difficult to predict. In this study the original approach [31] was, furthermore, extended from γ-proteobacteria to all species and extended to not only give a prediction, but also identify which genes predicted to be most significantly associated with (or important for) pathogenicity or non-pathogenicity. Thus, the method will not only provide a prediction of pathogenicity, but may also be useful for identifying novel putative pathogenicity genes, supporting further functional genomic studies.

The predictor has been implemented as a free to use web-service, called *PathogenFinder*, to which users can upload raw reads, obtained from different NGS sequencing platforms, as well as assembled genomes, and obtain a fast estimation of the pathogenic potential of the bacteria they are studying, as well as the identification of potentially pathogenic genes. *PathogenFinder* could be helpful in situations of possible bacterial outbreaks, in which a fast analysis of the unknown strain is important to save lives, and follows the direction modern clinical microbiology [32] and global epidemiology [33] are taking driven by the revolution brought by high throughput DNA sequencing technologies.

Results and Discussion

Overview on the Created Models

In this work we developed a method for predicting the pathogenicity of novel bacteria. We did this by comparing the proteins of the strain under investigation to a protein family database (PFDB) composed of groups of proteins (protein families or PFs) that were either associated with pathogenic or non-pathogenic organisms. In the creation of the PFDB we used 885 complete bacterial genomes (Table S1), 372 of which were tagged as human pathogens and 513 as non-pathogens.

All the proteins encoded by the bacterial genomes were initially clustered, and significant clusters, in which the majority of the proteins originated from either pathogens or non-pathogens, were identified. The PFs were accordingly tagged as pathogenic or non-pathogenic and a weight (Z-score) was calculated for each of them (see Materials and Methods for further details). Eight models were built using bacteria belonging to the same phylum or class as training data (Table 1). These models are named TM-*taxname*, where *taxname* is the phylum or class (e.g., bacteroidetes) of the organisms in the training data. Two other models created were: the whole-data model (WDM), which was trained using all the 885 bacteria in our training-set; the complement model (COMPL), which was trained using the organisms belonging to classes and/or phyla for which we had either only pathogenic or non-pathogenic strains and for which it was hence not possible to create specific models (Table S1).

Given a query organism, based on the number and kind of PFs that the proteins of the query organism are similar to, a prediction on whether it is human pathogenic or non-pathogenic is performed. The predictor has been implemented as a free to use web-server called *PathogenFinder*, to which a user can upload either the raw reads or the complete or draft genome of the organism they want to assess the pathogenicity of. One of the 10 built models can be selected for the prediction, and if the user does not know which class or phylum the organism belongs to, the web-server will identify it automatically by predicting 16S genes, using *RNAmmer* [34], and accordingly select the appropriate model to be used for the prediction.Both the set of matches used for the prediction and the raw matches from *PathogenFinder* are downloadable. The latter is particularly useful, since it contains more information about pathogenicity than the standard server output,

Table 5. Top 10 ranking pathogenic protein families and annotated functions of their proteins for the WDM model.

RANK	Z-score	P	N	Function
1	10.18	38	0	Borrelia Plasmid partition proteins
2	9.49	33	0	TCS associated genes, unknown functions
3	9.19	31	0	Lipoate-protein ligase, lipoate metabolism associated proteins
4	9.19	31	0	Unknown functions, flavin oxidoreductase
5	9.04	30	0	Exfoliative toxin A
6	8.89	29	0	Pili assembly proteins, Motility, Secretion Systems
7	8.89	30	0	Unknown function, shikimate kinase
8	8.89	29	0	Pili assembly proteins, Motility, Secretion Systems
9	8.74	28	0	Multiple antibiotic resistance (MarR) family proteins
10	8.74	28	0	Mutarotase Yjht (sialic acid mutarotation), unknown functions

P and N columns contain the number of pathogenic and non-pathogenic organisms in the protein family respectively.
doi:10.1371/journal.pone.0077302.t005

and could hence be used for a more detailed analysis of the pathogenicity features of the organisms under investigation.

Performance on Five-Fold Cross Validation and Independent Test Data

The TM models were tested using only organisms belonging to the specific phylum/class, while in the case of the WDM model the whole independent data-set was used for the test.

Table 2 shows the performance of the ten models as obtained by 5-fold cross validation (CV) (column 2) and on independent test-sets of organisms from the same taxonomic group (column 3). As can be seen for the Tenericutes and Bacteroidetes phyla, the performances were very poor when compared to the MCC obtained in the CV tests. This is likely to be caused by the models being built using a small number of organisms (Table 1). For instance, the TM-Tenericutes model was trained on only 14 isolates. Furthermore, it was tested on a set of organisms from species that were not present in the training-set.

To compare the performance of the WDM model to those of the TM and COMPL models, we examined the MCC of the WDM on the same test-sets used for the other models (column 4 in Table 2).

For example, to examine the performance of the WDM in predicting the pathogenicity of Firmicutes bacteria, we tested it with the same organisms used to assess the accuracy of the TM-Firmicutes model.

The MCC obtained by the WDM (0.758) on all bacteria was higher than the overall accuracy of all the TM models and COMPL model combined (0.736). Nonetheless, the TM models performed better for bacteroidetes, α, β, and γ-proteobacteria, even though for the latter the difference from the WDM was not significant. The remaining TM models and the COMPL model had lower MCC than the WDM for the same organisms.

Performance on Draft Genomes and *Escherichia coli*

The models ability in predicting the pathogenicity of an isolate as based on a draft genome was tested using 259 sets of illumina raw reads from 6 different species. While in the case of *Campylobacter jejuni*, *Klebsiella pneumoniae* and *Staphylococcus aureus* (57 isolates in total) all the predictions were correct, the results were not satisfactory for Enterococci and *E. coli*. Of 50 *Enterococcus*

Table 6. Top 10 ranking non-pathogenic protein families and annotated functions of their proteins for the WDM model.

RANK	Z-score	P	N	Function
1	−6.68	0	63	tRNA proteins
2	−6.62	0	62	ABC transporter related proteins (for δ and α-proteobacteria)
3	−6.18	0	54	Rubrerythrin
4	−6.07	0	52	Rubrerythrin
5	−6.01	0	51	Iron-sulfur binding domain proteins
6	−6.01	0	51	Hydroxymethylglutaryl-CoA synthase
7	−5.95	0	50	Unknown function
8	−5.89	0	49	Unknown function
9	−5.83	0	48	Unknown function
10	−5.70	0	46	Sulfite reductase subunit

P and N columns contain the number of pathogenic and non-pathogenic organisms in the protein family respectively.
doi:10.1371/journal.pone.0077302.t006

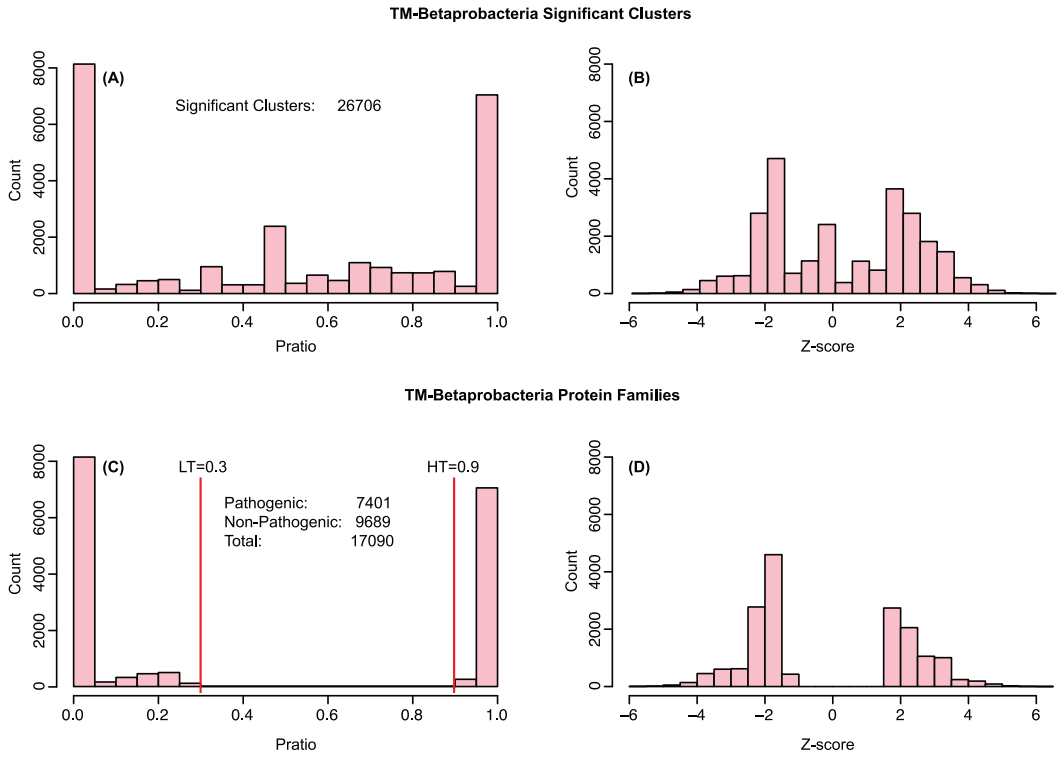


Figure 1. P_{ratio} and Z-score histograms for TM-Betaproteobacteria model. The model was built setting MinOrg = 2, HT = 0.9 and LT = 0.3. (A) and (B) respectively show the P_{ratio} and Z-score histograms for the clusters i such that $ORG_i \geq MinOrg$. By this step the original 69,744 clusters are reduced to 26,706. In (A) the bars at the extremes are the count for clusters containing either only genes from pathogenic organisms (right bar) and non-pathogenic ones (left bar), while the small pick in the middle are clusters containing the same number of pathogenic and non-pathogenic organisms, and hence will not be used since they provide no discriminative information about pathogenicity. (C) and (D) show the same histograms for the PFs obtained removing all the significant clusters with P_{ratio} value between LT and HT. We can see how the amount of non-pathogenic PFs is higher than the pathogenic ones (C). HT and LT can be used to modify the amount of both pathogenic and non-pathogenic PFs, which can be useful in model in which the training-set has an unbalanced amount of pathogenic and non-pathogenic organisms. In (D) the negative Z-scores are associated with non-pathogenic families while the others are for pathogenic PFs.
doi:10.1371/journal.pone.0077302.g001

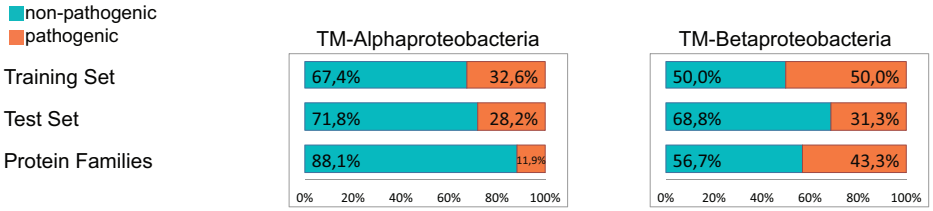
faecalis and 49 *Enterococcus faecium* from healthy Danish pigs, all isolates were predicted as pathogenic. Our training-set only contained a single pathogenic *E. faecalis* and no *E. faecium*, which may explain these results.

The WDM as well as the TM-Gammaproteobacteria models predicted the 10 *E. coli* strains in the test-set as pathogenic, although 4 strains were annotated as non-pathogenic. A similar situation was observed for the 103 *E. coli* draft genomes. Accordingly, we decided to create a model only for the *Enterobacteriaceae* family, using the organisms in our training-set. The resulting model correctly predicted 1 of 4 non-pathogenic *E. coli* achieving an MCC of 0.41, but all draft genomes were still predicted as pathogenic. The model also showed improvements in predicting other *Enterobacteriaceae*, with an MCC of 0.675, while WDM and TM-Gammaproteobacteria had an MCC of 0.519 and 0.617, respectively.

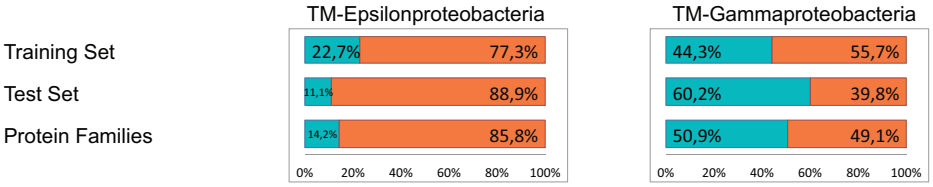
To improve the predictions for *E. coli* further, we decided to create 2 special models. These models were called *ecoli_boost* and *enterobac_boost*, and they were trained on a set that was enriched with 14 extra non-pathogenic *E. coli* strains downloaded from the National Center for Biotechnology Information (NCBI) (Table S2). These two models had a noticeably improvement on both CG test-sets and on the 103 assembled *E. coli* isolates, on which MCC was 0.346 (Acc = 67%) and 0.360 (Acc = 68%) for *enterobac_boost* and *ecoli_boost*, respectively. The lists of organisms used to train the *enterobac_boost* and *ecoli_boost* models, together with more details on the results on *E. coli* can be seen in Table S2.

Comparison to other Prediction Methods

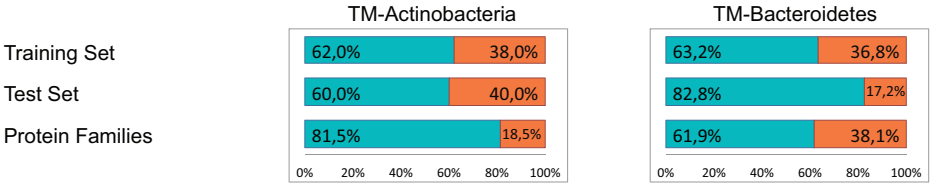
Presently, the literature describes two main approaches for predicting the human pathogenicity of bacteria based on whole genome sequencing data: the first, proposed by Andreatta et al. [31], is able to predict the pathogenicity of α-proteobacteria, and it was from this study we borrowed the concept of PFs; the second method, developed by Iraola et al. [28], uses SVM [35], and can predict the pathogenicity of all types of bacteria. In this method the authors selected 120 genes associated to pathogenicity from 600 complete genomes using SVM, and built a prediction model based on the selected genes.



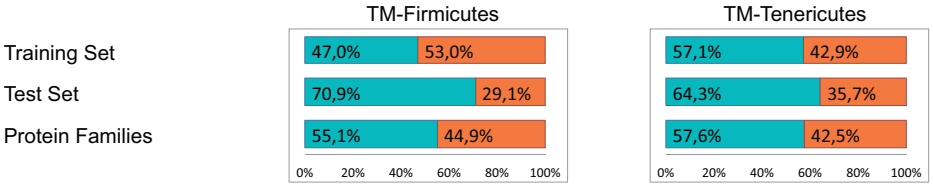
	TM-Alphaproteobacteria	TM-Betaproteobacteria
Tot. Protein Families	32961	17090
PFDB Size (No.Seq./Mbytes)	179809/82,7	76905/35,2



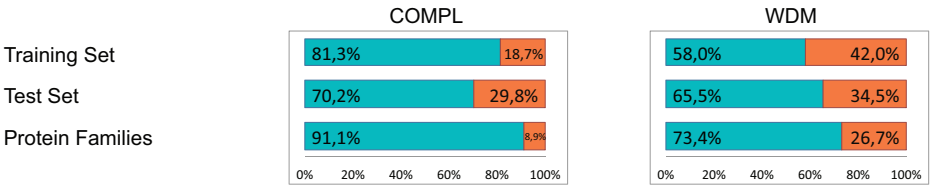
	TM-Epsilonproteobacteria	TM-Gammaproteobacteria
Tot. Protein Families	1494	40537
PFDB Size (No.Seq./Mbytes)	13128/4,4	229030/100,6



	TM-Actinobacteria	TM-Bacteroidetes
Tot. Protein Families	19018	9620
PFDB Size (No.Seq./Mbytes)	179809/29,6	33502/16,9



	TM-Firmicutes	TM-Tenericutes
Tot. Protein Families	16392	1536
PFDB Size (No.Seq./Mbytes)	99754/41,7	3965/2,0



	COMPL	WDM
Tot. Protein Families	67150	192411
PFDB Size (No.Seq./Mbytes)	311858/142,1	784316/346,5

Figure 2. PFDB, training and test-set for each model. Each bar-plot shows the percentage of pathogenic (orange) and non-pathogenic (light-blue) organisms in the training and test-set, and the percentage of pathogenic and non-pathogenic protein families in the PFDB of the model identified by the title of the bar-plot (eg. WMD). Below each horizontal bar-plot the number of protein families composing the PFDB of the model the bar-plot refers to, along with its size in megabytes and the number of sequences, is shown.
doi:10.1371/journal.pone.0077302.g002

To compare our method to the one proposed by Andreatta et al., we built a model using the same set of α-proteobacteria organisms (155) and the same parameters (MinOrg, HT, LT) used by Andreatta et al. The key differences between our method and the one by Andreatta et al. are: 1) we used CD-HIT instead of BLAST in both the protein clustering and prediction phases; 2) we used Equation 3 to filter the significant matches of the query sequences, while Andreatta et al. filtered based on a BLAST e-value threshold; 3) We compute the final predictions using the Z-scores, while Andreatta et al. counted the number of pathogenic and non-pathogenic families matched. The obtained model was tested on the same independent set used by Andreatta et al. This set included 24 organisms (14 pathogenic), and our model was able to correctly classify 23 organisms (95.8%). This is equivalent to an MCC of 0.92, while Andreatta's MCC was 0.837. The one organism that our method was not able to correctly classify, is *Salmonella enterica* Serovar Gallinarum str. 287/91 [GenBank:30689], which is pathogenic for poultry, but not known to be for humans. The pathogenicity of this organism is restricted to chicken although it shares a high quantity of genomic features associated to pathogenicity with its human pathogenic ancestor *Salmonella Enteriditis* [36]. It is likely that these features mislead the prediction model, since also the method by Andreatta et al. wrongly classified this *S. enterica* strain.

To compare our method to the predictor proposed by Iraola et al., we used the test-set they used for their blind test evaluation. The test-set, originally composed of 233 organisms, contained 5 strains, which were excluded from the comparison, since they were also present in our training-set. Overall, for the comparison, we had a test-set composed of 228 organisms, 192 of which are tagged as human pathogens and the remaining 36 as non-pathogens.

PathogenFinder achieved an overall MCC of 0.67 for the taxonomy models and 0.65 for the WDM model. Both results are higher than the MCC of 0.6 obtained by the method proposed by Iraola et al. Table S3 contains a detailed description of the comparison, including the organisms used and the corresponding predictions from both methods.

PFDB Analysis and Biological Interpretation

For each created model, an analysis of its PFDB was performed and its PFs ranked based on their Z-scores. The scores above 0 are associated with pathogenic PFs, while those below 0 are associated with non-pathogenic PFs. No protein function analysis was done prior to the models creation, making the approach unbiased on the genomic content of the organisms, regardless of their pathogenicity. In this paragraph we describe the analysis of the PFs of the TM-Gammaprobacteria and WDM.

The analysis of the PFDB of TM-Gammaproteobacteria model showed that the high ranked pathogenic families (Table 3) contained proteins well known to be associated to pathogenicity. The family at rank 1 and 3 contained N-acetylmannosamine kinase, which is a key enzyme in sialic acid synthesis and sialic acid transport proteins. Sialic acid is important for virulence and is believed to help the microbes to disguise themselves as host cells in order to elude the host's immune system response [37]. Fimbrial proteins (rank 2) are important for bacterial adherence [38]. At rank 10 we found cytochrome b_{562} proteins that help bacteria to survive and grow in conditions of poor oxygen [39]. Other high-ranked families contained proteins associated with secretion systems (II and III) and antibiotic resistance.

An interesting finding, which was also found in [31], was the presence of families containing proteins with unknown functions associated with pathogenicity. This finding suggests that those proteins with unknown function might have important roles in the bacterial pathogenesis and could form the basis for further functional studies improving our understanding of bacterial pathogenicity. Proteins with unknown functions were also identified as associated with non-pathogenic PFs (Table 4).

The analysis of the PDBF of the WDM enabled us to see if proteins involved (or not involved) in pathogenesis belong to organisms of different taxonomy, and at the same time gave us an insight on how proteins are conserved along the different phyla. Again, we found that the top ranked families associated to pathogenicity (Table 5) contained also proteins with unknown function.

The highest ranked PF contained proteins encoded by plasmids from different pathogenic *Borrelia* species (mainly *Burgdorferi*), which are involved in pathogenesis [40,41]. The family ranked 3rd contained proteins associated with lipoate metabolism. The acquisition and use of lipoate by pathogens affect their virulence and the pathogenesis of the diseases they cause [42]. Among the toxins found were: exofiliative toxin A (family-rank 5) in *Staphylococcus aureus* strains, causative of Staphylococcal scalded skin syndrome [43,44]; streptolysin (O and S), mainly found in *Streptococcus* pathogenic species [45]; hemolysin (II, III, α and β types) found in PFs mainly composed of α-proteobacteria [46,47] and firmicutes organisms [48]; shiga toxin, a common pathogenicity factor in many virulent *E. coli* strains [49]; dermonecrotic toxin (DNT), one of the main virulence factors in many *Bordetella* species [50](pertusiss in human), but at the same time present in plant pathogenic organisms like *Erwinia amylovora* [51] and *Erwinia pyrifoliae* [52]. The fact that we could find PFs containing DNT tagged as pathogens and others tagged as non-pathogenic (like the one containing DNT for *E. amylovora* and *E. pyrifoliae*) is an example of the ability of our clustering method to associate a given protein (a toxin in this case) to human pathogenicity as well as non-pathogenicity depending on the organism in which it is found.

Another example through which we could see the discriminative power of our PFs, was in associating pathogenicity to the different secretion system types proteins (SST1–SST6). For SST3 we identified 284 protein families, 147 of which were tagged as pathogenic. The pathogenic PFs were composed of human pathogenic α-proteobacteria strains, while the non pathogenic PFs contained plant pathogenic organisms from proteobacteria genera like *Xanthomonas*, *Agrobacter* and *Erwinia*, which use SST3 (and other secretion systems) to infect the hosts cells of plants [53,54].

The protein families with high rank associated with non-pathogenicity (Table 6) were usually composed of proteins present in bacteria living in hot springs, lake surfaces or deep in the sea, and the functions are associated to their ability to survive under those extreme environmental conditions. Among those proteins are Rubrerythrin, found in anaerobic sulphate-reducing bacteria like *Geobacter* and *Desulfivibrio* [55]. When the PFs were not composed of proteins from environmental bacteria, they contained mainly probiotics or plant pathogens. It is important to note that

since the WDM model was created with HT and LT parameters with values of 1.0 and 0 respectively, we only have PFs composed of proteins from either only pathogenic organisms or only non-pathogenic organisms.

Conclusions

There is an increasing need for fast identification of unknown bacteria with particular focus on the assessment of their potential pathogenicity. In this work we presented *PathogenFinder*, a web-server that by analysing the user-uploaded proteome can identify genomic features associated with both pathogenicity and non-pathogenicity. Given an input proteome the method quickly predicts its potential pathogenicity, making it a useful tool to be used together with other web-services developed for bacterial outbreak surveillance. Moreover, the possibility for the user to download the complete set of predicted pathogenicity features for the input organism makes *PathogenFinder* convenient for the analysis of pathogenic and harmless strains for microbiologists, epidemiologist and in general institutions studying bacterial pathogenesis.

One of the novel aspects in our approach is in the construction of the prediction models, which was carried out without any prior analysis of the proteins in our training-set, by just tagging our organisms as pathogenic or non-pathogenic and identifying protein families that were frequently found in pathogenic or non-pathogenic organisms.

It is important to notice that even though an isolate may have been obtained from a non-pathogenic environmental or animal or human related source it is not necessarily non-pathogenic. Such strain might in fact be highly pathogenic opportunistic pathogens. This naturally makes the creation of the optimal reference database difficult, but with increased number of isolates with well-defined meta data this is should still be doable.

We observed how *PathogenFinder* performs better than other pathogenicity prediction methods described in the literature, which usually rely on taxonomy and global sequence similarity with small sets of genes known to be associated with bacterial pathogenesis. We had less good results for species of the tenericutes phylum, and extra work need to be done to obtain statistically significant results for opportunistic strains (e.g. S. aureus) for which we could not tag any of our strains as non-pathogenic. The accuracy in predicting opportunistic bacteria could be improved by building specific models (e.g. at species level) as soon as new strains are available and there is a reasonable amount of both pathogenic and harmless strains. We have also shown how the prediction accuracy can be enhanced by increasing the number of organisms in the training-sets and/or making specific models at different taxonomic ranks, showing the example of *E. coli*, which is particularly difficult to predict because of the high similarity between commensal and pathogenic strains.

With the fast growing number of available bacterial complete genomes and with the increasing quality of the meta data we envision the possibility in the near future to build prediction models targeting only bacteria of a given genus or species, or even better, to build models to identify pathogenic features involved in specific diseases.

Materials and Methods

Training and Test Data

All available complete bacterial genomes (NCBI Genome Project, accessed on 10th Nov. 2010) were considered for the creation of the training-sets.

The pathogenicity information for the retrieved organisms were taken from NCBI genome project pages as described in Andreatta et al. [31], and for 885 of the 1,224 downloaded organisms, we were able to find pathogenicity information. The final complete training-set (Table S1) was composed of 513 organisms tagged as human non-pathogens and 372 tagged as human pathogens. For the human pathogenic organisms we checked for evidence in the literature.Opportunistic pathogens (e.g. from species like *Staphylococcus aureus* [56] or *Pseudomonas aeruginosa* [57]) were still tagged as pathogenic even though it has been shown that some of them can live inside the host without causing any disease, and their pathogenicity is sometimes related to the host's health conditions.

From January 2012, NCBI removed pathogenicity information from its pages, redirecting the users to Genomes Online Database (GOLD) [58]. On 26th Feb. 2012 we queried GOLD for pathogenicity information about organisms that had been published after 5th Nov. 2010 (the date of the latest published bacteria in the training-set). We were able to extract pathogenicity information for 449 organisms, and subsequently retrieved the corresponding complete genomes and plasmids from NCBI based on the NCBI project ids.

The final test data (Table S1) was composed of 449 organisms, 294 of which were tagged as human non-pathogens and 155 as human pathogens.

Protein Clustering

The model creation consisted of the following 2 main steps:

I. Protein Clustering
II. PFDB Creation

The initial idea for clustering the proteins was to use BLAST [59], but due to the size of our dataset (almost 3 million proteins), it would not have been computationally feasible. Instead, we used CD-HIT [60], which made it possible to cluster all the proteins in approximately 24 days using 2 3 Ghz dual-core CPUs in parallel and a 8 Gb of RAM.

The output from the program were 3 files containing respectively: 1) a list of cluster ids followed by the FASTA headers of the sequences composing the clusters; 2) a FASTA file containing all the clusters representative sequences; 3) a FASTA file containing all the solitary sequences that could not be included in any cluster.

Protein Family Database (PFDB) Creation

Our prediction models are based on the concept of protein families as initially proposed in Adreatta et al. [31]. Protein families are groups of proteins with a certain degree of similarity. The PFs were created using a two-steps filtering of the clusters created using CD-HIT. To perform this filtering we used four parameters: MinORG, P_{ratio}, LT and HT.

Let ORG be the number of organisms which have proteins in a given cluster *i*. We define MinORG as the minimum number of organisms that must have proteins in the *i* cluster for it to be considered significant. As such, MinORG is a lower threshold for the ORG value.

Equation 1. Ratio of human pathogenic organisms having proteins inside the cluster *i* on the total number of organisms having proteins in *i*. Newton's Second Law

$$P_{ratio}(i) = \frac{HP_i}{ORG_i} \quad (1)$$

P_{ratio} (Equation 1) is the ratio of the number of pathogen organisms having proteins in the *i* cluster (HP_i) on the total

number of organisms in i (ORG_i). LT and HT are thresholds for the P_{ratio} that we used to define if a given significant cluster should be tagged as pathogenic or non-pathogenic according to equation 2.

Equation 2. Function used to define if a given significant cluster should be tagged as 'pathogen family' or 'non-pathogen family'.

$$f(i)=\begin{cases} -1 & \text{if } P_{ratio}(i)\leq LT \\ 0 & \text{if } LT<P_{ratio}(i)<HT \\ 1 & \text{if } P_{ratio}(i)\geq HT \end{cases} \quad (2)$$

Let f (Equation 2) be the function we use to decide if a given significant cluster should be tagged as pathogenic or non-pathogenic. If the number of sequences from pathogens and non-pathogens is too close in a given cluster (if $P_{ratio}=0.5$ then $f(i)=0$), the cluster does not have any discriminative value for pathogenicity and is unusable.

Given a protein cluster i, it was considered a protein family if the following 3 conditions were satisfied:

I. $ORG_i \geq MinORG$

II. $f(i)\neq 0$

III. $P_{ratio}\geq HT$ or $P_{ratio}\leq LT$

The significance of a protein family depends on its ORG value and its P_{ratio}. A statistical measure called Z-score ($\mathcal{Z}$) was used to take into account the above two values of a family and assess its significance. The estimation of the $\mathcal{Z}$ values was performed on the set C composed of all the clusters i satisfying condition I. Let μ and σ be the average and standard deviation respectively of the P_{ratio} of the clusters in C. $\mathcal{Z}$ is a measure representing by how many standard deviations σ the mean x of a sample (a cluster in our case) differs from the mean μ of the population. Given a cluster i in C, its mean correspond to its P_{ratio} and we calculate the $\mathcal{Z}$ value for i as follows:

$$Z_i=\frac{P_{ratio}(i)-\mu}{SE_i}$$

Where SE is the standard error of the mean for i, and it is:

$$SE_i=\frac{\sigma}{ORG_i}$$

To each protein family, a $\mathcal{Z}$ value was assigned, and these are used in the calculation of the final prediction score as well as a ranking value in the analysis of the protein families. Figure 1 shows the distributions of the P_{ratio} values and Z-scores for both significant clusters and protein families for the TM-Betaproteobacteria model, while Figure 2 shows for each of the models built the proportion of pathogenic and non-pathogenic families in the PFDB, together with the training-set and test-set for the 10 models built. All the sequences in the PFDB are used to perform the predictions.

Models Optimisation

The prediction models were verified by 5-fold cross validation. For each of the models, many trials and tests were performed before choosing the MinOrg, LT and HT parameters for the final models. At each CV a parameter called $\mathcal{Z}thr$, was further optimised. $\mathcal{Z}thr$ is the threshold used to decide whether an input organism should be predicted as pathogenic or not, by comparing it to the summation of $\mathcal{Z}$ values obtained for the matching sequences in the input proteome. The parameters (MinOrg, HT, LT)(Table 1) of the models with the highest MCC in the CV tests were used to create the final models, and the corresponding $\mathcal{Z}thr$ values will be used as thresholds for the predictions.

Pathogenicity Prediction

The prediction method takes as input a FASTA file containing the proteins of the organism for which we want to assess the potential pathogenicity. In case the input is a complete or draft genome, initial gene prediction is performed using PRODIGAL [61]. PRODIGAL outputs a set of proteins representing the predicted genes. This is then used as input to our method. Using CD-HIT-2D [60], the input file is compared to the PFDB, and the output will contain all the input sequences that matched sequences in the PFDB, and that are used to compute the final prediction.

The following 4 steps describe the process that leads to the prediction:

I. Compare the input proteins to the PFDB

II. Filter hits based on the identity threshold (Equation 3)

III. Calculate final score summing the $\mathcal{Z}$ values associated to the matched PFs

IV. Compare the final score to the model's $\mathcal{Z}thr$ threshold and give the final prediction.

From the comparison in step I, we obtain a list of clusters, the representatives of which are sequences belonging to the PFDB, while the non-representative sequences come from the input. Because it is possible that more than one of the input proteins fall inside the same cluster, the sequence with the highest identity percentage with the representative is chosen. [!ht].

Equation 3. Calculates the identity threshold to select significant matches that will be used in the final prediction. The calculation is based on statistics on the identity values obtained for all matching query sequences.

$$idenThr(hits)=\begin{cases} \dfrac{\mu+\sigma+max}{2} & \text{if } \mu+\sigma\leq max \\ max & \text{otherwise} \end{cases} \quad (3)$$

The list of matches is then filtered based on an identity threshold that is dynamically computed at each prediction using the function idenThr (Equation 3). Let *hits* be a set containing all the percentage identity values for all our matches. Let μ and σ be respectively the average and standard deviation of the percentage identity values in *hits*. Let *max* be the maximum percentage identity obtained for the hits in PFDB. Remembering that, based on the settings of CD-HIT-2D, the minimum identity is 60%. Equation 3 calculates the identity threshold as the middle point between the maximum, and the average increased by one standard deviation, of the identities in *hits*. Selecting all the matches with an identity higher than idenThr(hits), we will obtain a list of hits with a very high identity relatively to the distribution of identities of our hits.

The matches below that threshold will not be used in the final prediction. The process will sometimes greatly reduce the number of matches, but this is in favour of matches with higher identity, making the final prediction more reliable, if compared to the

results obtained using a fixed threshold, as we proved by using the paired student's t-test (results not shown).

In the end we compute the summation of the Z-scores associated with the families matching the input sequences (III). If the sum of the Z-scores is above $\mathcal{Z}thr$ the input is considered pathogenic, otherwise it is considered non pathogenic (IV).

Supporting Information

Table S1 Training and Test organisms. xlsx file containing the list of organisms in the training and test-set and a table showing the phyla of the organisms in the training-set used to build the COMPL model.
(XLSX)

Table S2 Extra Escherichia Coli Strains. xlsx file containing the training-sets used for building *ecoli_boost* and *enterobac_boost* models, including the list of extra *E. coli* strains and a summary of the results in the prediction of *E. coli* and enterobacteriaceae organisms.
(XLSX)

Table S3 Comparison with other methods. xlsx file containing a detailed description of the comparison of *PathogenFinder* and the method described in [28].
(XLSX)

Acknowledgments

We are grateful to Federico De Masi for his help in the phase of submission of the manuscript.

Author Contributions

Conceived and designed the experiments: SC OL. Performed the experiments: SC. Analyzed the data: SC MVL. Contributed reagents/materials/analysis tools: MVL FMA. Wrote the paper: SC. Helped review the manuscript: MVL FMA OL.

References

1. WHO The global burden of disease: 2004 update [cited 2013 sep 13] http://www.who.int/healthinfo/global_burden_disease/GBD_report_2004update_full.pdf.
2. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, et al. (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464: 59–65.
3. Hooper LV, Gordon JI (2001) Commensal host-bacterial relationships in the gut. Science 292: 1115–1118.
4. Sassetti CM, Boyd DH, Rubin EJ (2003) Genes required for mycobacterial growth defined by high density mutagenesis. Molecular Microbiology 48: 77–84.
5. Young RA, Mehra V, Sweetser D, Buchanan T, Clark-Curtiss J, et al. (1985) Genes for the major protein antigens of the leprosy parasite mycobacterium leprae. Nature 316: 450–452.
6. Brogden KA, Guthmiller JM, Taylor CE (2005) Human polymicrobial infections. The Lancet 365: 253–255.
7. DuPont AW, DuPont HL (2011) The intestinal microbiota and chronic disorders of the gut. Nature Reviews Gastroenterology and Hepatology 8: 523–531.
8. Gogarten JP, Townsend JP (2005) Horizontal gene transfer, genome innovation and evolution. Nature Reviews Microbiology 3: 679–687.
9. Falkow S (1997) What is a pathogen. ASM news 63: 359–365.
10. Stock AM, Robinson VL, Goudreau PN (2000) Two-component signal transduction. Annual review of biochemistry 69: 183–215.
11. Hacker J, Kaper JB (2000) Pathogenicity islands and the evolution of microbes. Annual Reviews in Microbiology 54: 641–679.
12. Galán JE, Collmer A (1999) Type III secretion machines: Bacterial devices for protein delivery into host cells. Science 284: 1322–1328.
13. Gillespie JJ, Wattam AR, Cammer SA, Gabbard JL, Shukla MP, et al. (2011) PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. Infection and Immunity 79: 4286–4298.
14. Chen L, Xiong Z, Sun L, Yang J, Jin Q (2012) VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. Nucleic acids research 40: D641–645.
15. Frost LS, Leplae R, Summers AO, Toussaint A (2005) Mobile genetic elements: the agents of open source evolution. Nature Reviews Microbiology 3: 722–732.
16. Manning SD, Motiwala AS, Springman AC, Qi W, Lacher DW, et al. (2008) Variation in virulence among clades of escherichia coli O157:H7 associated with disease outbreaks. Proceedings of the National Academy of Sciences 105: 4868–4873.
17. Ho Sui SJ, Fedynak A, Hsiao WWL, Langille MGI, Brinkman FSL (2009) The association of virulence factors with genomic islands. PLoS ONE 4: e8094.
18. Wassenaar TM, Gaastra W (2001) Bacterial virulence: can we draw the line? FEMS Microbiology Letters 201: 1–7.
19. Paine K, Flower DR, et al. (2002) Bacterial bioinformatics: pathogenesis and the genome. Journal of molecular microbiology and biotechnology 4: 357–365.
20. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, et al. (1995) The minimal gene complement of mycoplasma genitalium. Science 270: 397–404.
21. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, et al. (1995) Whole-genome random sequencing and assembly of haemophilus inuenzae rd. Science 269: 496–512.
22. Fredericks DN, Relman DA (1996) Sequence-based identification of microbial pathogens: a reconsideration of koch's postulates. Clinical microbiology reviews 9: 18–33.
23. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376–380.
24. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456: 53–59.
25. Metzker ML (2010) Sequencing technologies – the next generation. Nature Reviews Genetics 11: 31–46.
26. Garg A, Gupta D (2008) VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens. BMC Bioinformatics 9: 62.
27. Nanni L, Lumini A, Gupta D, Garg A (2012) Identifying bacterial virulent proteins by fusing a set of classifiers based on variants of chou's pseudo amino acid composition and on evolutionary information. IEEE/ACM Trans Comput Biol Bioinformatics 9: 467–475.
28. Iraola G, Vazquez G, Spangenberg L, Naya H (2012) Reduced set of virulence genes allows high accuracy prediction of bacterial pathogenicity in humans. PloS one 7: e42144.
29. Day WA, Fernández RE, Maurelli AT (2001) Pathoadaptive mutations that enhance virulence: Genetic organization of the cadA regions ofShigella spp. Infection and Immunity 69: 7471–7480.
30. Maurelli AT (2007) Black holes, antivirulence genes, and gene inactivation in the evolution of bacterial pathogens. FEMS Microbiology Letters 267: 1–8.
31. Andreatta M, Nielsen M, Aarestrup F, Lund O (2010) In silico prediction of human pathogenicity in the -proteobacteria. PloS one 5: e13680.
32. Didelot X, Bowden R, Wilson DJ, Peto TEA, Crook DW (2012) Transforming clinical microbiology with bacterial genome sequencing. Nature Reviews Genetics 13: 601–612.
33. Aarestrup FM, Brown EW, Detter C, Gerner-Smidt P, Gilmour MW, et al. (2012) Integrating genome-based informatics to modernize global disease monitoring, information sharing, and response. Emerging Infectious Diseases 18: e1.
34. Lagesen K, Hallin P, Rødland EA, Stærfeldt HH, Rognes T, et al. (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Research 35: 3100–3108.
35. Cortes C, Vapnik V (1995) Support-vector networks. Machine learning 20: 273–297.
36. Thomson NR, Clayton DJ, Windhorst D, Vernikos G, Davidson S, et al. (2008) Comparative genome analysis of salmonella enteritidis PT4 and salmonella gallinarum 287/91 provides insights into evolutionary and host adaptation pathways. Genome research 18: 1624–1637.
37. Severi E, Hood D, Thomas G (2007) Sialic acid utilization by bacterial pathogens. Microbiology 153: 2817–2822.
38. Krogfelt KA (1991) Bacterial adhesion: genetics, biogenesis, and role in pathogenesis of fimbrial adhesins of escherichia coli. Review of Infectious Diseases 13: 721–735.
39. Turner SM, Moir JW, Griffiths L, Overton TW, Smith H, et al. (2005) Mutational and biochemical analysis of cytochrome c, a nitric oxide-binding lipoprotein important for adaptation of neisseria gonorrhoeae to oxygen-limited growth. Biochemical Journal 388: 545.
40. Stewart PE, Byram R, Grimm D, Tilly K, Rosa PA (2005) The plasmids of *borrelia burgdorferi*: essential genetic elements of a pathogen. Plasmid 53: 1–13.
41. Grimm D, Eggers CH, Caimano MJ, Tilly K, Stewart PE, et al. (2004) Experimental assessment of the roles of linear plasmids lp25 and lp28–1 of borrelia burgdorferi throughout the infectious cycle. Infection and immunity 72: 5938–5946.
42. Spalding MD, Prigge ST (2010) Lipoic acid metabolism in microbial pathogens. Microbiology and Molecular Biology Reviews 74: 200–228.
43. Lina G, Gillet Y, Vandenesch F, Jones ME, Floret D, et al. (1997) Toxin involvement in staphylococcal scalded skin syndrome. Clinical Infectious Diseases 25: 1369–1373.

44. Amagai M, Matsuyoshi N, Wang ZH, Andl C, Stanley JR (2000) Toxin in bullous impetigo and staphylococcal scalded-skin syndrome targets desmoglein 1. Nature Medicine 6: 1275–1277.
45. Alouf JE (1980) Streptococcal toxins (streptolysin o, streptolysin s, erythrogenic toxin). Pharmacology & therapeutics 11: 661–717.
46. Zhang XH, Austin B (2005) Haemolysins in vibrio species. Journal of Applied Microbiology 98: 1011–1019.
47. Chain PSG, Carniel E, Larimer FW, Lamerdin J, Stoutland PO, et al. (2004) Insights into the evolution of yersinia pestis through whole-genome comparison with yersinia pseudotuberculosis. Proceedings of the National Academy of Sciences of the United States of America 101: 13826–13831.
48. Ma S, ZhI B, Iv G, AIu T, Np K (1992) [evidence of the existence of hemolysin II from bacillus cereus: cloning the genetic determinant of hemolysin II]. Molekuliarnaia biologiia 27: 1218–1229.
49. Paton JC, Paton AW (1998) Pathogenesis and diagnosis of shiga toxin-producing escherichia coli infections. Clinical Microbiology Reviews 11: 450–479.
50. Walker KE, Weiss AA (1994) Characterization of the dermonecrotic toxin in members of the genus bordetella. Infection and Immunity 62: 3817–3828.
51. Metzger M, Bellemann P, Bugert P, Geider K (1994) Genetics of galactose metabolism of erwinia amylovora and its inuence on polysaccharide synthesis and virulence of the fire blight pathogen. Journal of Bacteriology 176: 450–459.
52. Smits THM, Jaenicke S, Rezzonico F, Kamber T, Goesmann A, et al. (2010) Complete genome sequence of the fire blight pathogen erwinia pyrifoliae DSM 12163T and comparative genomic insights into plant pathogenicity. BMC genomics 11: 2.
53. Jones JDG, Dangl JL (2006) The plant immune system. Nature 444: 323–329.
54. Abramovitch RB, Anderson JC, Martin GB (2006) Bacterial elicitation and evasion of plant innate immunity. Nature Reviews Molecular Cell Biology 7: 601–611.
55. Moura I, Tavares P, Ravi N (1994) [15] characterization of three proteins containing multiple iron sites: Rubrerythrin, desulfoferrodoxin, and a protein containing a six-iron cluster. In: Harry D Peck J, editor, Methods in Enzymology, Academic Press, volume Volume 243. 216–240. URL http://www.sciencedirect.com/science/article/pii/0076687994430179.
56. den Heijer CD, van Bijnen EM, Paget WJ, Pringle M, Goossens H, et al. (2013) Prevalence and resistance of commensal *staphylococcus aureus*, including meticillin-resistant *s aureus*, in nine european countries: a cross-sectional study. The Lancet infectious diseases.
57. Stover C, Pham X, Erwin A, Mizoguchi S, Warrener P, et al. (2000) Complete genome sequence of pseudomonas aeruginosa pao1, an opportunistic pathogen. Nature 406: 959–964.
58. Pagani I, Liolios K, Jansson J, Chen IMA, Smirnova T, et al. (2011) The genomes OnLine database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Research 40: D571–D579.
59. Altschul S, Madden T, Schäffer A, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research 25: 3389–3402.
60. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22: 1658–1659.
61. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, et al. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11: 119.

3 BACTERIAL SPECIES TYPING AND IDENTIFICATION

Identification and classification is a basic requirement for anyone working with bacteria, whether commensals, pathogens, or bacteria used for industrial purposes. When the bacterium under investigation is a new or mutated strain that could potentially cause an outbreak, it is of crucial importance that the identification of the microbe causing the infection is rapid and precise.

Knowing what bacteria it is, usually means knowing what species the microbe belongs to. For the last 30 years, taxonomy identification of prokaryotes has mainly been done through the analysis of 16S rRNA gene sequences. The 16S-based method has, however, some limitations and other methods for taxonomic classification of prokaryotes using WGS data has been proposed.

In the first manuscript presented in this chapter we trained and benchmarked five methods for bacterial species identification on common datasets of complete and draft genomes, and raw-reads. One of the methods (SpeciesFinder) is based on the complete 16S rRNA gene, while the others use different bioinformatics approaches, like KmerFinder, which analyses the co-occurrence of k-mers in the input DNA sequences to identify the bacterial species. The results of the benchmark show how methods that only use core genes in the prediction have difficulties in identifying species, which have recently diverged. On the other hand, the method based on k-mers had overall the highest accuracy and robustness.

The second article included in this chapter [40] is about accurate strain identification (or typing). We developed a web-server that runs a bioinformatics method for bacterial strain typing based on multilocus sequence typing (MLST). The web-service uses a database of MLST schemes from *www.pubmlst.org*, to which it is weekly synchronised. Having a web-server like the MLST available for free reduces both the costs and time needed for scientists that want to perform strain identification.

The traditional MLST, performed in the lab, costs about 120$ per isolate, while strain-typing using pulsed-field gel electrophoresis (PFGE) costs approximately 150$ per isolate [63]. Considering that the WGS for a single bacterial isolate can be obtained for less than 100$ and the MLST web-service can be used free of charge, the advantages in terms of costs are clear, which might be particularly important for small microbiology labs in developing countries.

The MLST web-service is at present used by researchers from more than 40 countries worldwide, and serves more than 700 bacterial strain identifications every month.

Manuscript II

Multilocus Sequence Typing of
Total-Genome-Sequenced Bacteria

Journal of Clinical Microbiology

Multilocus Sequence Typing of Total-Genome-Sequenced Bacteria

Mette V. Larsen, Salvatore Cosentino, Simon Rasmussen, Carsten Friis, Henrik Hasman, Rasmus Lykke Marvig, Lars Jelsbak, Thomas Sicheritz-Pontén, David W. Ussery, Frank M. Aarestrup and Ole Lund

J. Clin. Microbiol. **2012, 50(4):1355. DOI: 10.1128/JCM.06094-11.**
Published Ahead of Print 11 January 2012.

JCM
Journals.ASM.org

Multilocus Sequence Typing of Total-Genome-Sequenced Bacteria

Mette V. Larsen,[a] Salvatore Cosentino,[a] Simon Rasmussen,[a] Carsten Friis,[b] Henrik Hasman,[b] Rasmus Lykke Marvig,[c] Lars Jelsbak,[c] Thomas Sicheritz-Pontén,[a] David W. Ussery,[a] Frank M. Aarestrup,[b] and Ole Lund[a]

Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark[a]; National Food Institute, Technical University of Denmark, Lyngby, Denmark[b]; and Center for Systems Microbiology, Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark[c]

Accurate strain identification is essential for anyone working with bacteria. For many species, multilocus sequence typing (MLST) is considered the "gold standard" of typing, but it is traditionally performed in an expensive and time-consuming manner. As the costs of whole-genome sequencing (WGS) continue to decline, it becomes increasingly available to scientists and routine diagnostic laboratories. Currently, the cost is below that of traditional MLST. The new challenges will be how to extract the relevant information from the large amount of data so as to allow for comparison over time and between laboratories. Ideally, this information should also allow for comparison to historical data. We developed a Web-based method for MLST of 66 bacterial species based on WGS data. As input, the method uses short sequence reads from four sequencing platforms or preassembled genomes. Updates from the MLST databases are downloaded monthly, and the best-matching MLST alleles of the specified MLST scheme are found using a BLAST-based ranking method. The sequence type is then determined by the combination of alleles identified. The method was tested on preassembled genomes from 336 isolates covering 56 MLST schemes, on short sequence reads from 387 isolates covering 10 schemes, and on a small test set of short sequence reads from 29 isolates for which the sequence type had been determined by traditional methods. The method presented here enables investigators to determine the sequence types of their isolates on the basis of WGS data. This method is publicly available at www.cbs.dtu.dk/services/MLST.

Correct, standardized classification is a basic need for anyone working with bacteria, whether pathogens, commensals, or bacteria used for industrial purposes. Especially in outbreak situations, it is of pivotal importance that the strains of infectious agents be rapidly and accurately identified. A recent example is the outbreak of hemolytic-uremic syndrome and bloody diarrhea caused by an *Escherichia coli* O104:H4 strain, which in the beginning of May 2011 started spreading in Germany and throughout Europe. Reliable classification, including determination of the multilocus sequence type (MLST), was needed to identify strains related to the outbreak (19, 23). Also, for a range of other species, MLST is used to classify isolates in an understandable and comparable global context (6, 12, 22, 31).

MLST was first developed for *Neisseria meningitidis* in 1998 to overcome the poor reproducibility between laboratories of older molecular typing schemes (18). The principle behind the MLST scheme is to identify internal nucleotide sequences of approximately 400 to 500 bp in multiple housekeeping genes. Unique sequences (alleles) are assigned a random integer number, and a unique combination of alleles at each locus, an "allelic profile," specifies the sequence type (ST). Following the introduction of the *Neisseria* MLST scheme, MLST has been considered the "gold standard" of typing, and additional schemes that cover bacterial and fungal species have been developed. The MLST allele sequences and ST profile tables are stored in curated databases hosted at different sites around the world (1, 14, 15). The PubMLST site collects data from all databases and makes it easily accessible (multilocus sequence typing databases and software, December 2011 [http://pubmlst.org]).

Traditionally, MLST starts with a PCR amplification step using primers that are specific for the loci of the MLST scheme, followed by Sanger sequencing. The procedure is both costly and time-consuming. In this new era of high-throughput sequencing, it may be more rational to use whole-genome sequence (WGS) data for typing. The cost of DNA sequencing has steadily gone down roughly 10-fold every 5 years (25), and the development of next- and third-generation sequencing methods has provided equally great reductions in equipment investments, thus making the technology accessible to individual investigators and routine clinical and microbial laboratories. The challenge, however, is to extract the relevant information from the large amount of data generated by these techniques. To allow comparison with results obtained by other commonly used technologies and with historical data, it is also important to be able to relate the WGS data to typing schemes such as MLST.

We present here the publicly available MLST server (www.cbs.dtu.dk/services/MLST), which uses WGS data for identifying the STs of bacteria.

MATERIALS AND METHODS

MLST databases. MLST allele sequences and ST profile tables are stored in online databases hosted at five different sites around the world. The University of Oxford collects data from all databases and makes it easily accessible (http://pubmlst.org). In total, 66 bacterial MLST schemes are currently available. Most of them function at the species level, e.g., *Escherichia coli* and *Staphylococcus aureus* schemes, while a few function on the genus level, e.g., the *Bifidobacterium* and *Neisseria* schemes. Most schemes include 7 housekeeping genes, but schemes with as few as 5 and as many as 10 genes have also been developed. For four bacterial species, two different MLST schemes are available: *Acinetobacter baumannii* (2; Institut Pasteur,

Received 12 October 2011 Returned for modification 28 November 2011
Accepted 29 December 2011
Published ahead of print 11 January 2012
Address correspondence to Mette V. Larsen, metteb@cbs.dtu.dk.
Supplemental material for this article may be found at http://jcm.asm.org/.

doi:10.1128/JCM.06094-11
The authors have paid a fee to allow immediate free access to this article.

Acinetobacter baumannii MLST database, December 2011 [http://www.pasteur.fr/recherche/genopole/PF8/mlst/Abaumannii.html]), *Clostridium difficile* (10, 17), *E. coli* (13, 32), and *Pasteurella multocida* (28; PubMLST, *Pasteurella multocida* multi-host MLST databases, December 2011 [http://pubmlst.org/pmultocida_multihost/]).

Data sets. (i) Assembled genomes. In August 2010, 1,212 completely sequenced and assembled bacterial genomes were collected from the NCBI Genome database (http://www.ncbi.nlm.nih.gov/sites/genome). For 336 of these genomes, MLST schemes have been developed and are available through the MLST databases (Table 1).

(ii) Sequence reads. Table 2 shows an overview of the species for which we had short sequence reads, along with the sequencing platforms used. *Campylobacter jejuni*, *E. coli*, *Klebsiella pneumoniae*, *Staphylococcus aureus*, and *Salmonella enterica* isolates were sequenced on the Illumina platform generating paired-end reads by TGen (United States). *Streptococcus thermophilus*, *Bifidobacterium animalis*, and *Lactococcus lactis* isolates were sequenced on the Illumina platform generating paired-end reads by Source BioScience (United Kingdom), BaseClear (The Netherlands), and BGI (Hong Kong). Other *S. thermophilus*, *B. animalis*, *B. longum*, and *L. lactis* isolates were sequenced on the Illumina platform generating single reads by Source BioScience (United Kingdom). *P. aeruginosa* isolates were sequenced on the Illumina platform generating single reads by Partners HealthCare (Boston, MA) or on the Roche 454 GS platform by the Allegheny-Singer Research Institute (Pittsburgh, PA). WGS data for the 2011 German O104:H4 *E. coli* outbreak were obtained from publicly available sources. Data from 7 × 314 chips sequenced on the Ion Torrent platform (Life Technologies) for a single isolate were obtained from BGI (23). Data from 8 × 314 chips sequenced on the Ion Torrent platform (Life Technologies) for the outbreak strain LB226692 were obtained from Life Technologies and the University of Münster (19). Illumina MiSeq single-read data for five different isolates of the outbreak strain were obtained from the British Health Protection Agency (HPA). From the Göttingen Genomics Laboratory, we obtained data for two isolates of the outbreak strain sequenced on a Roche 454 GS sequencer. The ABI SOLiD data were the 50 × 50 mate pair data set with 600× coverage of *E. coli* DH10B, available from the SOLiD software development community.

Draft assembly of short sequence reads. If sequence reads are given as input to the MLST server, the reads are assembled *de novo* prior to ST prediction. Short-read data produced from all major next- and third-generation sequencing platforms, such as the Illumina, Roche 454 GS, and Applied Biosystems SOLiD platforms and the Life Science Ion Torrent personal genome machine (PGM), are supported (8, 18a, 24, 26, 30). The *de novo* assembly creates contiguous sequences without gaps from the DNA sequence reads, termed contigs, and when paired-end or mate-paired reads are available, these are used to combine the contigs into scaffolds. As a measure of the quality of the draft assembly, the N_{50} value is calculated for the assembled genomes. The N_{50} value for contigs or scaffolds is defined as the length of the shortest contig or scaffold in the set of the largest contigs or scaffolds that represents at least 50% of the assembly (20). The assembly is available for download from the MLST server.

Illumina sequence data are assembled using Velvet, version 1.1.04 (34). Prior to assembly, paired-end data are filtered and trimmed using the following steps. (i) All reads containing the character N are removed. (ii) If a read matches at least 15 nucleotides (nt) of a sequencing primer/adaptor, the read is trimmed at the 5′ coordinate of the match. (iii) The 3′ tail is trimmed up to a quality score of 15 (phred scale). (iv) The minimum average quality of the read after trimming is 20. (v) The length of the read after trimming is at least 15 nt. We do not trim Illumina single-end data, since benchmarking showed that this reduced the overall quality of the assemblies and of MLST prediction for the data sets used in the study. Then, in parallel, several assemblies using *k*-mer sizes from 33% to 80% of the average read length are run, and the assembly with the best cumulative rank for N_{50}, number of contigs, and length of the largest contig is selected as the best assembly.

Both Roche 454 GS and Ion Torrent PGM sequence data are assembled using the Roche proprietary GS De Novo Assembler software, version 2.6 (Newbler 2.6). If given standard flowgram files (.sff), the assembler clips and trims the data prior to assembly.

For Applied Biosystems SOLiD sequence data, assembly is performed using the SOLiD System *de novo* Accessory Tools, version 2.2. The assembly pipeline uses colorspace Velvet 0.7.55 (34) for the assembly and is run without read error correction (with SAET) or postassembly analysis in order to decrease run time. For all sequencing technologies, single-end, paired-end, or mate-paired reads can be used for assembly.

After uploading of short read data, the assembly is available for download from the MLST server.

Implementation of MLST on completely sequenced bacteria. An automatic weekly download script was set up for all allele sequences and ST profiles from the MLST databases. Via a script written in Perl, the assembled bacterial genome was converted into a BLAST database. Using the specified MLST scheme, the genome was searched by BLAST for all MLST alleles for all genes. Statistically significant alignments between the query sequence (the MLST alleles) and sequences in the BLAST genome database are called high-scoring segment pairs (HSP) according to BLAST terminology. As the Expect threshold, we use the default value, which is 10.

The best-matching MLST allele is found by calculating the length score (LS) as QL − HL + *G*, where QL is the length of the MLST allele, HL is the length of the HSP, and *G* is the number of gaps in the HSP. The allele with the lowest LS and, secondly, with the highest percentage of identity (ID) is selected as the best-matching MLST allele. A perfectly matching MLST allele will have an LS of zero and 100% ID, meaning that all the nucleotides of the MLST allele match with the nucleotides in the genome across the entire length of the allele. Note that the BLAST HSP E value or score cannot be used for selecting the correct allele, since a long allele with a percentage of ID below 100% can have a lower E value (i.e., a higher score) than a shorter allele with 100% ID. Per definition, the shorter allele with 100% ID over the whole length is the correct allele.

After identification of the MLST allele for all genes of the MLST scheme, the ST is determined on the basis of the combination of identified alleles.

RESULTS

MLST implementation. For MLST of completely sequenced bacterial genomes, short sequence reads are, in a first step, assembled to draft genomes as described in Materials and Methods. It is also possible to bypass the assembly step and to input a complete or partial preassembled genome. The minimum requirement for a partial genome is that it contain all the loci necessary for MLST. For a specific MLST scheme, the MLST alleles of each locus are aligned to the genome by using BLAST. The closest-matching MLST allele is selected, and the ST is determined based on the combination of MLST alleles. Two different output formats are available. The short output format includes the identified ST and details about the concordance of each locus with the best-matching MLST allele in the database. Figure 1 shows an example of the short output format from the typing of a *P. aeruginosa* isolate. The extended output format additionally includes the nucleotide sequences of the MLST alleles identified (see Fig. S1 in the supplemental material). This format can be useful for drawing phylogenetic trees.

MLST of 336 assembled bacterial genomes. To evaluate our method, we used it for identification of the STs of 336 completely sequenced and preassembled bacterial genomes. These bacteria cover 56 MLST schemes. Table 1 shows the results with regard to the proportion of the MLST alleles in the tested genomes that were previously unseen and hence were not registered in the MLST

TABLE 1 MLST of preassembled, completely sequenced bacterial isolates

MLST scheme	No. of loci in scheme	Avg no. of alleles per locus[a]	No. of STs[a]	No. of isolates[b]	Proportion[c] of:	
					New alleles	Unknown STs
*Acinetobacter baumannii*_1	7	82	346	6	0.095	0.333
*Acinetobacter baumannii*_2	7	34	124	6	0.000	0.167
Arcobacter	7	205	357	2	0.214	0.500
Bacillus cereus	7	129	553	10	0.043	0.100
Bifidobacterium	7	42	102	11	0.221	0.273
Bordetella	7	8	43	5	0.400	0.400
Borrelia burgdorferi	8	125	402	2	0.000	0.000
Brachyspira	7	39	36	3	0.571	1.000
Brachyspira hyodysenteriae	7	17	66	1	0.143	0.000
Burkholderia pseudomallei	7	46	886	4	0.000	0.000
Corynebacterium diphtheriae	7	40	227	1	0.000	0.000
Campylobacter fetus	7	10	35	1	0.000	0.000
Campylobacter jejuni	7	415	5,489	6	0.000	0.000
Campylobacter lari	7	50	18	1	0.000	0.000
Campylobacter upsaliensis	7	42	138	2	0.000	0.500
Clostridium botulinum	7	10	24	11	0.377	0.455
*Clostridium difficile*_1	7	18	128	2	0.000	0.000
*Clostridium difficile*_2	7	14	65	2	0.000	0.000
Cronobacter	7	48	74	2	0.000	0.000
Enterococcus faecalis	7	61	435	1	0.000	0.000
Enterococcus faecium	7	48	617	26	0.011	0.038
*Escherichia coli*_1	7	228	2,333	36	0.004	0.000
*Escherichia coli*_2	8	143	535	36	0.031	0.250
Flavobacterium psychrophilum	7	15	33	1	0.000	0.000
Haemophilus influenzae	7	124	939	4	0.071	0.000
Haemophilus parasuis	7	25	116	1	0.000	0.000
Helicobacter pylori	7	2,088	2,356	10	0.543	0.600
Klebsiella pneumoniae	7	94	688	3	0.000	0.000
Lactobacillus casei	7	9	40	3	0.000	0.333
Leptospira	7	25	117	6	0.667	0.667
Listeria monocytogenes	7	79	34	6	0.000	0.000
Mannheimia haemolytica	7	13	35	3	0.000	0.000
Moraxella catarrhalis	8	40	214	1	0.000	0.000
Neisseria	7	561	8,999	8	0.000	0.000
Pasteurella multocida multihost	7	25	46	1	0.000	0.000
Pasteurella multocida RIRDC	7	47	189	1	0.000	0.000
Porphyromonas gingivalis	7	32	138	2	0.000	0.000
Propionibacterium acnes	7	12	58	2	0.000	0.000
Pseudomonas aeruginosa	7	116	1,070	4	0.036	0.250
Stenotrophomonas maltophilia	7	47	56	2	0.000	0.000
Salmonella enterica	7	395	1,492	18	0.008	0.167
Sinorhizobium	10	18	136	2	0.000	0.000
Staphylococcus aureus	7	244	2,107	21	0.000	0.000
Staphylococcus epidermidis	7	34	361	2	0.000	0.000
Streptococcus agalactiae	7	58	557	3	0.000	0.000
Streptococcus pneumoniae	7	319	6,947	12	0.012	0.000
Streptococcus pyogenes	7	89	572	13	0.022	0.000
Streptococcus suis	7	87	239	6	0.238	0.500
Streptococcus thermophilus	6	22	116	3	0.111	0.000
Streptococcus uberis	7	42	475	1	0.000	0.000
Streptomyces	6	107	135	5	0.733	0.600
Vibrio parahaemolyticus	7	141	348	1	0.000	0.000
Vibrio vulnificus	10	40	83	2	0.400	1.000
Wolbachia	5	168	236	4	0.000	0.000
Xylella fastidiosa	7	17	27	4	0.000	0.000
Yersinia pseudotuberculosis	7	11	95	4	0.000	0.000

[a] Registered in the MLST database.
[b] Number of isolates with completely sequenced genomes tested for this scheme.
[c] Proportion of alleles found in the isolates, or proportion of STs found for the isolates, which were not already registered in the database.

TABLE 2 MLST of completely sequenced bacterial isolates using short sequence reads

Sequencing platform	Species	No. of isolates	MLST scheme	Proportion of loci with:		Avg N_{50}	Log avg N_{50}
				Minor mismatches	Major mismatches		
Illumina							
Paired-end reads	*B. animalis*	5	*Bifidobacterium*	0.000	0.029	33,113	4.52
	C. jejuni	53	*Campylobacter*	0.005	0.000	131,571	5.12
	E. coli	15	*E. coli* scheme 1	0.010	0.000	195,822	5.29
	E. coli	15	*E. coli* scheme 2	0.075	0.000	196,463	5.29
	K. pneumoniae	4	*K. pneumoniae*	0.000	0.000	207,167	5.32
	L. lactis	34	*L. lactis*	0.564	0.039	44,525	4.65
	S. aureus	83	*S. aureus*	0.017	0.009	196,736	5.29
	S. enterica	50	*S. enterica*	0.000	0.009	249,501	5.40
	S. thermophilus	13	*S. thermophilus*	0.090	0.000	47,686	4.68
Single reads	*E. coli*	6	*E. coli* scheme 1	0.000	0.000	46,479	4.67
	B. animalis	2	*Bifidobacterium*	0.000	0.071	24,979	4.40
	B. longum	2	*Bifidobacterium*	0.000	0.000	22,548	4.35
	L. lactis	7	*L. lactis*	0.238	0.119	14,114	4.15
	P. aeruginosa	81	*P. aeruginosa*	0.019	0.125	9,380	3.97
	S. thermophilus	9	*S. thermophilus*	0.000	0.000	28,823	4.46
Roche 454	*E. coli*	3	*E. coli* scheme 1	0.000	0.000	92,131	4.96
	P. aeruginosa	2	*P. aeruginosa*	0.000	0.000	60,477	4.78
Ion Torrent	*E. coli*	2	*E. coli* scheme 1	0.000	0.500	13,779	4.14
SOLiD	*E. coli*	1	*E. coli* scheme 1	0.286	0.000	165,835	5.22

databases. For 34 MLST schemes, all alleles in the MLST loci in the tested genomes matched perfectly to an allele already registered in the MLST databases (the proportion of new alleles equaled zero), while for the remaining 22 MLST schemes, 0.4% to 73.3% of the MLST alleles in the genomes were not in the MLST databases.

Two MLST schemes exist for *E. coli*: *E. coli* scheme 1, which employs seven genes (*adk*, *fumC*, *gyrB*, *icd*, *mdh*, *purA*, *recA*) (32), and *E. coli* scheme 2, which employs eight genes (*dinB*, *icdA*, *pabB*, *polB*, *putP*, *trpA*, *trpB*, *uidA*) (13). When the 36 completely sequenced *E. coli* isolates were typed using *E. coli* scheme 1, only one allele (0.4%) was not in the database. When *E. coli* scheme 2 was used, 10 alleles (3.1%) were not in the database. This difference in the proportion of previously unseen alleles may reflect either the coverage of the MLST databases (that is, how large a fraction of the total number of alleles they contain) or the rates of evolution of the genes used by the two schemes. The database for *E. coli* scheme 1 contains on average 228 alleles per locus, while that for *E. coli* scheme 2 contains on average 143 alleles per locus. Accordingly, the higher number of previously unseen alleles found by using *E. coli* scheme 2 seems mostly to reflect the fact that this database is less complete than the database for scheme 1.

The three MLST schemes that resulted in the highest number of previously unseen MLST alleles were the *Brachyspira* (57.1% new alleles), *Leptospira* (66.7% new alleles), and *Streptomyces* (73.3% new alleles) schemes. These schemes are meant to cover a whole genus rather than a specific species. As a consequence, these databases are expected to contain far more alleles than databases that aim at covering only a single species. However, this was not the case, as can be seen from Table 1. The *Neisseria* scheme also aims at covering a whole genus, but here no new alleles were found in the eight *Neisseria* genomes tested (two *Neisseria gonorrhoeae* and six *Neisseria meningitidis* genomes). Indeed, the *Neisseria* database is the second largest database, containing 561 alleles per locus, in accord with the early establishment of the database in 1998 (18). Of interest, the *Helicobacter pylori* database contains an average of 2,088 alleles per locus and as such is by far the largest database. Apparently, this does not mean that the database is in any way complete, since more than half of the alleles in the 10 *H. pylori* genomes tested are not in the database. This observation indicates that the genes selected for the *H. pylori* MLST scheme (33) are evolving faster than the genes that are generally used in the MLST schemes. This idea is in line with studies showing that in general, *H. pylori* has high rates of recombination and mutation (5, 7, 29).

Eight bacterial MLST schemes were not tested in this analysis, since we did not have access to complete genomes from these species. However, it is possible to use the MLST Web server with these species as well (*Brachyspira intermedia*, *Burkholderia cepacia* complex, *Campylobacter helveticus*, *Campylobacter insulaenigrae*, *Streptococcus oralis*, *Streptococcus equi* subsp. *zooepidemicus*, *Clostridium septicum*, and *Chlamydiales* spp.).

From short sequence reads to MLST. MLST implementation was then tested on short sequence reads from 387 bacterial isolates covering 10 MLST schemes and four sequencing platforms. Table 2 shows the results. We have divided the genomic MLST loci that did not perfectly match an MLST allele in the databases into major and minor mismatches. The major mismatches occur when the MLST allele from the MLST database exceeds the length of the contig, meaning that the MLST locus is only partly contained in the contig. In Fig. S1 in the supplemental material, the *aro* gene represents a major mismatch in a *P. aeruginosa* genome. The minor mismatches are all other types of mismatches and are equivalent to the "new alleles" of Table 1. In Fig. S1, the *acs* gene represents a minor mismatch.

MLST Results

Sequence Type: *Unknown ST**

**Please note that one or more loci do not match perfectly to any previously registered MLST allele. We recommend verifying the results by traditional methods for MLST.*

SETTINGS:

Organism: *Pseudomonas aeruginosa*

MLST Profile: *paeruginosa*

Genes in MLST Profile: *7*

Locus	%Identity	HSP Length / Allele Length	Gaps	Allele
acs	99.74%	390/390	0	*acs_28*
aro	100%	349/498	0	*aro_122*
gua	100%	373/373	0	*gua_11*
mut	100%	442/442	0	*mut_11*
nuo	100%	366/366	0	*nuo_4*
pps	100%	370/370	0	*pps_12*
trp	100%	443/443	0	*trp_3*

extended output

CONTIGS INFO:

Technology: *Illumina Single End Reads*

N50: *2670*

FIG 1 MLST results for a *P. aeruginosa* isolate in the short output format. By use of the MLST Web server, a *P. aeruginosa* strain that had been sequenced on the Illumina platform generating single reads was typed. For the purpose of the example, we have chosen to show the results obtained by using short sequence reads that assemble into a draft genome with a low N_{50}. Shown are the name of the loci in the MLST scheme, the percentage of nucleotides that are identical in the best-matching MLST allele in the database and the corresponding sequence in the genome (% identity), the length of the alignment between the best-matching MLST allele in the database and the corresponding sequence in the genome (also called the high-scoring segment pair [HSP]), the length of the best-matching MLST allele in the database, the number of gaps in the HSP, and the name of the best-matching MSLT allele. Note that for a perfectly matching allele, the percentage of identity will be 100%, the allele length will equal the HSP length, and the number of gaps will be zero. Green indicates a perfect match, while red indicates an imperfect match.

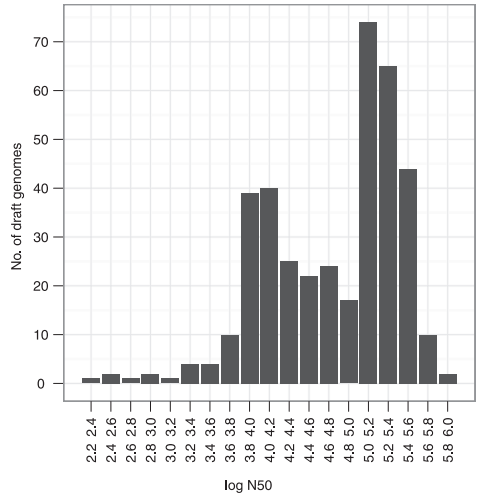


FIG 2 Distribution of log N_{50} values for 387 assembled draft genomes.

For 11 of the 15 sets of isolates sequenced by the Illumina technology for paired-end or single reads, and for all of the isolates sequenced on the Roche 454 GS platform, the frequency of alleles with minor mismatches was below 2%. For the remaining four sets of isolates, where the frequency of minor mismatches was above 2% (*E. coli*, Illumina paired-end reads, *E. coli* MLST scheme 2; *L. lactis*, Illumina paired-end and single reads; *S. thermophilus*, Illumina paired-end reads), this is likely to reflect the small size of the MLST database.

Whereas the proportion of alleles with minor mismatches reflects the coverage of the database for the selected scheme, the proportion of alleles with major mismatches reflects how well the short sequence reads have been assembled into a draft genome. The N_{50} value is a measure of the quality of the draft assembly: the higher the N_{50} value, the better the quality of the assembly. In general, Illumina paired-end reads were assembled into draft genomes with higher N_{50} values (average N_{50}, 165,149; 95% confidence interval [95% CI], 150,491 to 179,807) than Illumina single reads (average N_{50}, 13,943; 95% CI, 11,824 to 16,062). For the remaining sequencing platforms, we have too little data to draw conclusions on the general quality of the assembled draft genomes. Furthermore, the variability can be very large, as evidenced by the two *E. coli* isolates that were sequenced on the Life Sciences Ion Torrent PGM platform. While the isolate sequenced by Life Technologies and the University of Münster had an N_{50} value of 28,537, the isolate sequenced by BGI was assembled into a draft genome with an N_{50} of 666. As a comment on this poor N_{50} value, it should be noted that only the FASTQ files from the sequencing, not the flowgram files, were available to us.

For the assembled *P. aeruginosa* genomes, 13.2% of the alleles contained major mismatches. However, more than 40% of the alleles with major mismatches were found in the assembled genomes of only five isolates. The average N_{50} of these five draft genomes was as low as 503 (95% CI, 175 to 831).

Figure 2 shows the distribution of the log N_{50} for all assembled draft genomes. Fifteen draft genomes had a log N_{50} below 3.6 (N_{50} below 4,000). The remaining draft assemblies are contained within two peaks, roughly separating the draft genomes based on single reads from those based on paired-end reads.

For a small subset of the *P. aeruginosa* and *S. aureus* isolates, and for all *K. pneumoniae* isolates, the ST had been determined previously by traditional methods. For 10 of the *E. coli* isolates, the WGS data were obtained from publicly available sources. These isolates were all from the 2011 German *E. coli* O104:H4 outbreak, the causative agent of which has been found to belong to ST-678 (4, 19, 23). Table 3 shows that 25 of the 29 isolates with known STs were assigned the correct ST on the basis of our method for MLST. Three of the *P. aeruginosa* isolates were not assigned the correct ST. Instead, they all contained major mismatches and were assigned the ST "unknown" (N_{50} values, 371, 453, and 1,154). For the *E. coli* isolate sequenced by BGI using the Life Sciences Ion Torrent PGM, the MLST loci likewise contained major mismatches and the ST "unknown" was assigned.

DISCUSSION

WGS of bacterial pathogens has become an option for more scientists than formerly and even for routine laboratories due to the

TABLE 3 Isolates with known STs

Sequencing platform	Species	No. of isolates with known STs[a]	No. of correctly identified STs[b]
Illumina			
Paired-end reads	*S. aureus*	6	6
	K. pneumoniae	4	4
Single reads	*E. coli*	6	6
	P. aeruginosa	7	4
Roche 454	*E. coli*	2	2
	P. aeruginosa	2	2
Ion Torrent	*E. coli*	2	1

[a] Determined by traditional methods.
[b] Predicted by using WGS data.

declining costs of sequencing and the increasing number of analytic methods available. WGS may be useful in trend studies, in diagnostics, and for surveillance. Depending on the technology, WGS can be performed in a couple of hours. By combining this speed with low costs and the right tools, real-time surveillance and quick detection of outbreaks will become possible. As both the costs of technology and the run times continue to decline, WGS will become increasingly available to routine diagnostic laboratories. The challenges will thus be not to produce the sequence data but to extract the relevant information so as to allow for comparisons over time and between laboratories. Ideally, this information should also allow for comparison to historical data.

We have developed, implemented, and evaluated an MLST predictor based on WGS data. The method is publicly available at www.cbs.dtu.dk/services/MLST. The user can upload either a preassembled complete or partial bacterial genome or short sequence reads from one of four sequencing platforms. Currently, 70 different MLST schemes for 66 species are available.

The MLST Web server was specifically designed for ease of use, for the benefit of investigators with limited bioinformatics experience. The first step is to upload the preassembled genome or short sequence reads. In the case of short sequence reads, the sequence platform also needs to be specified. After one selects the MLST scheme to be used, the job can be submitted.

Jolley and Maiden have developed a Web-accessible database system, BIGSdb, that can also use WGS for MLST (16). This system, however, works only on UNIX/Linux systems and requires the installation of a whole range of programs and databases. The MLST Web server presented here can be used by anyone with a computer and a reasonably fast Internet connection.

Although new typing methods are expected to emerge in the wake of complete genome sequencing, e.g., single nucleotide polymorphism (SNP) typing (9, 11) and pangenome family trees (27), these methods lack standardized implementation and general acceptance in the scientific community. We therefore believe that MLST will still be considered the "gold standard" for typing for some time. In addition, for many years, knowledge of the ST will be crucial for comparison to data from isolates that were characterized before complete genome data became easily available.

The MLST server will continue to be improved, e.g., by addition of an option for the automatic detection of species, and hence the selection of the MLST scheme to be used, based on 16S rRNA typing. Furthermore, it will become possible to obtain a phylogenetic tree as output, which will enable the user to see how the ST of the query isolate relates to other STs.

Additional features for analyzing WGS data are also under development. These include the identification of antimicrobial resistance and virulence genes, as in a study described recently (3). Furthermore, we are developing methods for species identification and phylogenetic analysis based on SNP and pangenome analysis.

ACKNOWLEDGMENTS

This work was supported by the Center for Genomic Epidemiology at the Technical University of Denmark and was funded by grant 09-067103/DSF from the Danish Council for Strategic Research.

We are grateful to Mads Bennedsen, Birgitte Stuer-Lauridsen, and colleagues at Chr Hansen A/S (Hørsholm, Denmark) for sharing unpublished genome sequence data. We are grateful to Hans-Henrik Stærfeldt and John Damm Sørensen for excellent technical assistance.

REFERENCES

1. **Aanensen DM, Spratt BG.** 2005. The multilocus sequence typing network: mlst.net. Nucleic Acids Res. **33**:W728–W733.
2. **Bartual SG, et al.** 2005. Development of a multilocus sequence typing scheme for characterization of clinical isolates of *Acinetobacter baumannii.* J. Clin. Microbiol. **43**:4382–4390.
3. **Bennedsen M, Stuer-Lauridsen B, Danielsen M, Johansen E.** 2011. Screening for antimicrobial resistance genes and virulence factors via genome sequencing. Appl. Environ. Microbiol. **77**:2785–2787.
4. **Bielaszewska M, et al.** 2011. Characterisation of the *Escherichia coli* strain associated with an outbreak of haemolytic uraemic syndrome in Germany, 2011: a microbiological study. Lancet Infect. Dis. **11**:671–676.
5. **Bjorkholm B, et al.** 2001. Mutation frequency and biological cost of antibiotic resistance in *Helicobacter pylori.* Proc. Natl. Acad. Sci. U. S. A. **98**:14607–14612.
6. **David MD, Kearns AM, Gossain S, Ganner M, Holmes A.** 2006. Community-associated meticillin-resistant *Staphylococcus aureus*: nosocomial transmission in a neonatal unit. J. Hosp. Infect. **64**:244–250.
7. **Falush D, et al.** 2001. Recombination and mutation during long-term gastric colonization by *Helicobacter pylori*: estimates of clock rates, recombination size, and minimal age. Proc. Natl. Acad. Sci. U. S. A. **98**:15056–15061.
8. **Fedurco M, Romieu A, Williams S, Lawrence I, Turcatti G.** 2006. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Res. **34**:e22.
9. **Gardy JL, et al.** 2011. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N. Engl. J. Med. **364**:730–739.
10. **Griffiths D, et al.** 2010. Multilocus sequence typing of *Clostridium difficile.* J. Clin. Microbiol. **48**:770–778.
11. **Hendriksen RS, et al.** 2011. Population genetics of *Vibrio cholerae* from Nepal in 2010: evidence on the origin of the Haitian outbreak. mBio **2**(4): e00157–11. doi:10.1128/mBio.00157-11.
12. **Heym B, Le Moal M, Armand-Lefevre L, Nicolas-Chanoine MH.** 2002. Multilocus sequence typing (MLST) shows that the 'Iberian' clone of methicillin-resistant *Staphylococcus aureus* has spread to France and acquired reduced susceptibility to teicoplanin. J. Antimicrob. Chemother. **50**:323–329.
13. **Jaureguy F, et al.** 2008. Phylogenetic and genomic diversity of human bacteremic *Escherichia coli* strains. BMC Genomics **9**:560.
14. **Jolley KA, Chan MS, Maiden MC.** 2004. mlstdbNet—distributed multilocus sequence typing (MLST) databases. BMC Bioinformatics **5**:86.
15. **Jolley KA, Maiden MC.** 2006. AgdbNet—antigen sequence database software for bacterial typing. BMC Bioinformatics **7**:314.
16. **Jolley KA, Maiden MC.** 2010. BIGSdb: scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics **11**:595.
17. **Lemee L, Dhalluin A, Pestel-Caron M, Lemeland JF, Pons JL.** 2004. Multilocus sequence typing analysis of human and animal *Clostridium difficile* isolates of various toxigenic types. J. Clin. Microbiol. **42**:2609–2617.
18. **Maiden MC, et al.** 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc. Natl. Acad. Sci. U. S. A. **95**:3140–3145.

18a. **Margulies M, et al.** 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature **437**:376–380.
19. **Mellmann A, et al.** 2011. Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. PLoS One **6**:e22751.
20. **Miller JR, Koren S, Sutton G.** 2010. Assembly algorithms for next-generation sequencing data. Genomics **95**:315–327.
21. Reference deleted.
22. **Oliveira DC, Tomasz A, de Lencastre H.** 2002. Secrets of success of a human pathogen: molecular evolution of pandemic clones of meticillin-resistant *Staphylococcus aureus.* Lancet Infect. Dis. **2**:180–189.
23. **Rohde H, et al.** 2011. Open-source genomic analysis of Shiga-toxin-producing *E. coli* O104:H4. N. Engl. J. Med. **365**:718–724.
24. **Rothberg JM, et al.** 2011. An integrated semiconductor device enabling non-optical genome sequencing. Nature **475**:348–352.
25. **Service RF.** 2006. Gene sequencing. The race for the $1000 genome. Science **311**:1544–1546.
26. **Shendure J, et al.** 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. Science **309**:1728–1732.
27. **Snipen L, Ussery DW.** 2010. Standard operating procedure for computing pangenome trees. Stand. Genomic Sci. **2**:135–141.
28. **Subaaharan S, Blackall LL, Blackall PJ.** 2010. Development of a multilocus sequence typing scheme for avian isolates of *Pasteurella multocida.* Vet. Microbiol. **141**:354–361.
29. **Suerbaum S, et al.** 1998. Free recombination within *Helicobacter pylori.* Proc. Natl. Acad. Sci. U. S. A. **95**:12619–12624.
30. **Turcatti G, Romieu A, Fedurco M, Tairi AP.** 2008. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. Nucleic Acids Res. **36**:e25.
31. **Wagenlehner FM, et al.** 2007. Management of a large healthcare-associated outbreak of Panton-Valentine leucocidin-positive meticillin-resistant *Staphylococcus aureus* in Germany. J. Hosp. Infect. **67**:114–120.
32. **Wirth T, et al.** 2006. Sex and virulence in *Escherichia coli*: an evolutionary perspective. Mol. Microbiol. **60**:1136–1151.
33. **Wirth T, et al.** 2004. Distinguishing human ethnic groups by means of sequences from *Helicobacter pylori*: lessons from Ladakh. Proc. Natl. Acad. Sci. U. S. A. **101**:4746–4751.
34. **Zerbino DR, Birney E.** 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. **18**:821–829.

Manuscript III

Benchmarking of Methods for Genomic Taxonomy

Benchmarking of Methods for Genomic Taxonomy.

Mette Voldby Larsen [1], Salvatore Cosentino [1], Oksana Lukjancenko [1], Dhany Saputra [1], Simon Rasmussen [1], Henrik Hasman [2], Thomas Sicheritz Pontén [1], Frank M. Aarestrup [2], David Wayne Ussery [1,3] and Ole Lund [1]

[1] Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, 2800 Kongens Lyngby, Denmark
[2] National Food Institute, Technical University of Denmark, 2800 Kongens Lyngby, Denmark
[3] Comparative Genomics Group, Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA

Abstract

One of the first questions that emerge when encountering a prokaryotic organism of interest is what it is that is which species it is. The 16S rRNA gene formed the basis of the first method for sequence-based taxonomy and has had a tremendous impact on the field of microbiology. Nevertheless, the method has been found to have a number of shortcomings.

In the current study we trained and benchmarked five methods for whole genome sequence based prokaryotic species identification on a common dataset of complete genomes; 1) SpeciesFinder, which is based on the complete 16S rRNA gene, 2) Reads2Type that searches for species-specific 50-mers in either the 16S rRNA gene, the GyrB gene (for the Enterobacteraceae family) or the ITS gene (for the Mycobacterium genus), 3) The rMLST method that samples up to 53 ribosomal genes, 4) TaxonomyFinder, which is based on species-specific functional protein domain profiles, and finally 5) KmerFinder, which examines the number of co-occuring k-mers. The performances of the methods were subsequently evaluated on three datasets of short sequence reads or draft genomes from public databases. In total, the evaluation sets constituted more than 11,000 isolates covering 159 genera and 243 species. Our results indicate that methods that only sample chromosomal, core genes have difficulties in distinguishing closely related strains, which only recently diverged. The KmerFinder method had the overall highest accuracy and identified from 93%-97% of the isolates in the evaluations sets correctly to the species level.

Importance : The 16S rRNA locus has served as the backbone of prokaryotic taxonomy for more than 30 years, but has been recognized to be less than optimal for a number of species. The current advent of whole genome sequencing provides the opportunity to surpass 16S rRNA typing by including a larger fraction of the genome. Meanwhile, the amble amounts of WGS data in public databases enable us to perform educated proposals on how to optimally use this type of data.

INTRODUCTION

Rapid identification ofi solated bacterial species is essential for surveillance for human and animal health and for choosing the optimal treatment and control measures. Since the be-

Corresponding author, e-mail: mette@cbs.dtu.dk

ginning of microbiology more than a century ago, this has to a large extent been based on morphology and biochemical testing. However, for more than 30 years, 16S rRNA sequence data has served as the backbone for the classification of prokaryotes (1) and tremendous amounts of 16S rRNA sequences are available in public repositories (2; 3; 4). However, due to the conserved nature of the 16S rRNA gene, the resolution is often too low to adequately resolve different species and sometimes not even adequate for genus delineation (5; 6). Furthermore, many prokaryotic genomes contain several copies of the 16S rRNA gene with substantial inter-gene variation (7; 8). It is also considered problematic that this gene represents only a tiny fraction, roughly about 0.1% or less, of the coding part of a microbial genome (9).

Second- and third generation sequencing techniques have the potential to revolutionize the classification and characterization of prokaryotes. However, so far no consensus on how to utilize the vast amount ofi nformation in Whole Genome Sequence (WGS) data has emerged. Nevertheless, a number of different methods have been proposed. Roughly, they can be divided into those that require annotation of genes in the data and those that employ the nucleotide sequences directly.

One of the first attempts to employ WGS data for taxonomic purposes was carried out in 1999 (10). At the time, 13 completely sequenced genomes of unicellular organisms were available and distance-based phylogeny was constructed on the basis of presence and absence of suspected orthologous (direct common ancestry) gene pairs. Later it was recognized that methods that take into account gene content can be greatly influenced by Horizontal Gene Transfer (HGT) and alternative methods were developed that used homologous groups (gene family content) (11) or protein domains (12).

Functional protein domains also form the basis of a recent approach developed by our group (13). Here, the protein domains are combined into functional profiles of which some are species-specific and can thus be used for inferring taxonomy.

As an extension of 16S rRNA analysis, which focuses on a single locus, Super Multilocus Sequence Typing (SuperMLST) has been proposed (14). It relies on the selection of a set of genes that are highly conserved and hence can be used with any organism. In a publication from 2012, Jolley et al. suggested that 53 genes encoding ribosomal proteins are used for bacterial classification in an approach called ribosomal MLST (rMLST) (15). Not all 53 genes were found in all bacterial genomes, but due to the relatively high number of sampled loci, this is not considered as problematic. The rMLST method forms the basis of a proposed reclassification of Neisseria species (16) and has also been used for analyzing human Campylobacter isolates (17).

It is also possible to employ the sequence data directly without pre-annotation of genes. This can, for instance, be done by looking at k-mers (substrings of k nucleotides in DNA sequence data) that are sufficiently long to avoid co-occurrence in two random genomes. As an example, there are more than 4 billion different possible 16-mers, making their co-occurrence in two unrelated bacterial genomes unlikely. The number of co-occurring k-mers in two bacterial genomes can thus be considered a measure of evolutionary relatedness, and used to construct a phylogeny. Using this approach, all regions of the genome are considered, not only core genes. Furthermore, a gene segment will score highly despite the transposition of a gene segment within the genome, since only the flanking regions will be mismatched.

In the current study we have trained five different methods for species identification on a common dataset of complete prokaryotic genomes. 1) SpeciesFinder serves as the baseline, as it is based solely upon the 16S rRNA gene. 2) Reads2Type is a variant hereof, searching for species-specific 50-mers, predominantly within the 16S rRNA gene, with the help of non-species-specific 50-mers to quickly narrow down the search. 3) rMLST, which predicts species by examining 53 ribosomal genes. 4) TaxonomyFinder, which is based on species-specific functional protein domain profiles, and finally 5) KmerFinder, which predicts species

by examining the number of overlapping 16-mers.

The public available databases contain ample amounts of WGS data from prokaryotes, enabling us to conducting a large-scale benchmark study of the proposed methods. Hence, the process of reaching a consensus on how the WGS data should optimally be used for prokaryotic taxonomy is initiated.

MATERIALS & METHODS

Dataset

Training Data

In August 2011 a total of 1,647 complete genomes originating from Bacteria (1,535) and Archaea (112) were downloaded from the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/genome). For each genome, the annotated taxonomy according to GenBank was compared to the taxonomy according to Entrez, which was retrieved using the taxonomy module of BioPerl. Discrepancies were checked and corrected manually. For each genome, it was also examined if the annotated name was in accordance to the List of Prokaryotic names with Standing in Nomenclature (http://www.bacterio.cict.fr/allnames.html). When possible, names that were not in accordance were corrected to valid ones. In this way, 1,426 genomes were assigned to 847 approved genus and species names. The remaining 221 genomes, which were either only assigned to a genus, e.g., Vibrio spp., or assigned to species with informal names, e.g., Synechoccus islandicus, were left in the training data under the assumption that they will influence the different methods for species identification equally. An overview of the training data is available in Supplementary Table 1.

Evaluation Data

Three datasets were generated for the purpose of evaluating the methods. The first consisted of assembled complete of draft genomes with assigned species, which were downloaded from NCBI in September 2012 and not already part of the training data. Only genomes assigned to species that were also present in the training data were included. The set is called NCBI $_{drafts}$ and consists of genomes from 695 isolates covering 81 genera and 149 species. The set includes three Archaea; two Methanobrevibacter smithii and one Sulfolobus solfataricus. An overview of the data can be seen in Supplementary Table 2.

Furthermore, In January 2012, 11,768 sets of Illumina raw reads were downloaded from the NCBI Sequence Reads Archive (SRA, http://www.ncbi.nlm.nih.gov/sra) with assigned species (18). 10,517 of them had been sequenced by the Illumina Genome Analyzer II sequencer, while the remaining 1,251 had been sequenced by the Illumina HiSeq 2000 sequencer. Reads that could not be assembled to a draft genome were removed as were reads from species that were not present in the training. The final SRA $_{reads}$ dataset consists of 8,798 sets of paired-end reads and 1,609 sets of single reads, 10,407 sets in total.

The short reads of the SRA $_{reads}$ set were de novo assembled using velvet 1.1.04 (19). For of the draft assemblies the optimal k-mer length was estimated and used as described previously (20). The resulting set of draft genomes constitutes the SRA $_{drafts}$ evaluation set. To measure the qualities of the draft assemblies, the N50 values were calculated (21). The draft assemblies had an average N50 of 77,018, ranging from 101 to 779,945 (see Supplementary Figure 1), an average number of scaffolds of 697, and an average size of 3,301

kilobases. The SRA $_{reads}$ and SRA $_{drafts}$ sets both cover 167 different species from 120 genera with more than 5,000 strains from the Streptococcus, Staphylococcus and Salmonella genera. There are no species from Archeae. An overview of the SRA $_{reads}$ and SRA $_{drafts}$ sets is available in Supplementary Table 3.

Methods for species identification

SpeciesFinder

SpeciesFinder predicts the prokaryotic species based on the 16S rRNA gene. A 16S database was built from the genomes of the common training data using RNAmmer (22). The species predictions were performed differently depending on the input type. If the input was short reads, the prediction was done as follows:

I The reads were mapped against the 16S database using the Burrows-Wheeler aligner (BWA)(23).

II The BWA output was assembled using Trinity (24) to obtain the 16S rRNA sequences.

III The BLAST algorithm (25) was used to search the output from Trinity against the 16S database.

IV The best BLAST hit (see below) was chosen and the species associated with the best hit was given as the final prediction.

When the input sequence was a draft or complete genome, the prediction was performed as follows:

I The 16S rRNA gene was predicted from the input sequences using RNAmmer.

II Using the BLAST algorithm, the predicted sequence was aligned against the 16S database.

III The best BLAST hit (see below) was chosen and the species associated with it given as the final prediction.

The best BLAST hit was chosen by ranking the output from the BLAST alignment by a combination of coverage, percent identity, bitscore, number of mismatches, and number of gaps. The highest ranked hit was chosen for the prediction.
SpeciesFinder is available at http://cge.cbs.dtu.dk/services/SpeciesFinder/ .

rMLST

The rMLST method predicts bacterial species based on 53 ribosomal genes originally defined by Jolley et al. (15). The set of genes can either be used in an approach similar to Multilocus Sequence Typing (MLST), where each locus in the query genome is considered identical or non-identical to alleles of the corresponding locus in the reference database, and an allelic profile based on random numbers assigned to each of the alleles in the database is generated accordingly. Since the strains that we compare are more diverse than the ones compared in MLST, it is likely that many loci would have no identical matches in the database, making a simple cluster analysis based on allelic profiles problematic. To improve the resolution of the method, in our implementation of rMLST, the nucleotide sequence of each locus is aligned to the alleles in the reference database and a measure of the similarity of the locus and the

best matching allele is used subsequently, as described below.

Briefly, for each of the genomes in the training data, the 53 ribosomal genes were provided by Keith Jolley, Department of Zoology, University of Oxford, UK. In this way, for each genome, a gene collection of up to 53 ribosomal genes was assigned. To predict the species of a query genome, the query genome was first aligned to each gene collection using BLAT (26). Only hits with at least 95% identity and 95% coverage were considered as a potential match. If there were several potential matches, the best match was selected based on the best cumulative rank of coverage, percent identity, bitscore, number of mistmatches, and number of gaps in the alignments. The final prediction was given as the organism with the highest number of best hits across all genes. Our implementation of rMLST performs predictions for draft or complete genomes, but not short reads.

TaxonomyFinder

The TaxonomyFinder method is based on taxonomy group-specific protein profiles (ref). It performs predictions for draft or complete genomes, but not for short reads. The common training data was used to create the taxonomy-specific profile database. Briefly, for each genome functional profiles were assigned based on three collections of Hidden Markov Models (HMMs) databases: PfamA (27), TIGRFAM (28), and Superfamily (29). Genes that did not match any entry in the HMM databases were clustered using CD-HIT (30). Further, genomes were grouped according to the taxonomy level, either phylum or species, and profiles that were specific to each taxonomic group were extracted. Profiles were considered specific to a taxonomic group, if they were conserved in most of the genomes within a phylum/species group and absent in all genomes outside of the group. The workflow of the TaxonomyFinder method is a four-step process, which includes:

I Open-reading frame prediction using Prodigal (31).

II Construction off unctional profiles from protein-coding sequences.

III Assignment off unctional profiles.

IV Functional profile comparison to the taxonomy-specific profile database. The number of architectures, matched to each of the taxonomy groups, is recorded, and the fraction of taxa-specific genes (score) is calculated. The best-matching taxonomy group is selected based on a consensus of the best score and highest number of matched architectures.

TaxonomyFinder is available at http://cge.cbs.dtu.dk/services/TaxonomyFinder/ .

KmerFinder

The KmerFinder method predicts prokaryotic species based on the number of overlapping (co-occuring) k-mers, i.e. 16-mers between the query genome and genomes in a reference database. Initially, all genomes in the common training data were split into overlapping 16-mers with step-size one, meaning that if the first 16-mer is initiated at position N and ends at position N+15, the next 16-mer is initiated at position N+1 and ends at position N+16, and so on. To reduce the size of the final 16-mer database only 16-mers with the prefix ATGAC were kept. These 16-mers were stored in a hash table with links to the original genomes. When performing the prediction, the species of the query genome is predicted to be identical to the species of the genome in the training data with which

it has the highest number of 16-mers in common regardless of position. The input for KmerFinder can be draft or complete genomes as well as short reads. KmerFinder is available at http://www.cbs.dtu.dk/services/KmerFinder/ .

Reads2Type

Reads2Type identified the prokaryotic species based on a database of 50-mer probes generated from chosen marker genes (Saputra D., Rasmussen S., Larsen M.V., Haddad N., Aarestrup F.M., Lund O., and Sicheritz-Pontén T., submitted for publication). The version of Reads2Type evaluated in this study requires short reads as input. For bacterial species not belonging to the Enterobacteriaceae family or the Mycobacterium genus, the 50-mer database relies on the 16S rRNA locus, while for Enterobacteriaceae , the gyrB locus is used, and for Mycobacterium the ITS locus. Briefly, the following steps were applied for building the 50-mer probe database:

I 16S rRNA sequences of the complete bacterial genomes of the common training set were predicted using RNammer (22).

II For species belonging to the Enterobacteriaceae family or the Mycobacterium genus, gyrB sequences and ITS sequences, respectively, were downloaded from NCBI.

III The above sequences were pooled and all possible 50-bp fragments were generated from that pool.

IV 16S rRNA probes unique for Enterobacteriaceae and Mycobacteria were removed from the pool of 50-mers.

V All 50-mer duplicates associated to the conserved regions of different strains but the same species were removed.

VI To further reduce the size of the final 50-mers database, 25 consecutive 50-mers previously fragmented from one 50 bp stretch of 16s rRNA belonging to the same list of organism were removed.

The resulting 50-mers probe database consists of a number of sequences found uniquely in one species, as well as other sequences shared between several species. Subsequently, each read was compressed into a suffix tree, which is a data structure for fast string matching. The compressed short reads were aligned to the 50-mer probe database using a "narrow-down approach" strategy, i.e. when a compressed read matched a probe belonging to a group of species, a much smaller probe database excluding other species was created on the fly, causing the read progress to be faster and the species to be identified much faster.

The Reads2Type method is available as a web server (http://cge.cbs.dtu.dk/services/Reads2Type/) and as a console. The web-based Reads2Type is unique in not requiring the short read file to be uploaded to the server. Instead, the 4.6 MB 50-mers probe database is automatically transferred into the client computers memory before initiating the species identification. All computations needed for the species identification is fully performed on the clients computer, minimizing the data transfer and avoiding the network bottleneck on the server.

Testing the speed

The speed of the methods was evaluated on non-published internal data from up to 450 strains covering eight species (Enterococcus faecalis, Enterococcus faecium, Eschericia coli, Escherichia fergusonii, Klebsiella pneumoniaea, Salmonella enterica, Staphylococcus aureus, and Vibrio cholera) that had been sequenced by the Illumina sequencing method. Draft genomes were de novo assembled as described above for the SRA_{drafts} set. The speed was tested on a Cluster with x86_64 architecture, 128 nodes, 4 tasks per node, 30 or 7G per node.

RESULTS

Performances on NCBI draft genomes

The SpeciesFinder, rMLST, TaxonomyFinder, and KmerFinder methods are able to perform species predictions on draft or completed prokaryotic genomes. Their performances were evaluated on the NCBI $_{drafts}$ set of 695 draft genomes covering 149 species. Supplementary File 1 lists all predictions, while Figure 1A summarizes the results. Overall, SpeciesFinder, which is based on the 16S rRNA gene, had the poorest performance, only correctly identifying 76% of the isolates down to species level. KmerFinder, which is based on co-occurring 16-mers, had the highest performance and correctly identified 93% of the isolates. For only three isolates (0.43%), KmerFinder did not even get the genus correct. These three isolates were two Escherichia coli predicted as Shigella sonnei and one Providencia alcalifaciens predicted as Yersinia pestis.

The NCBI $_{drafts}$ set contains three Archaeal isolates; two M. smithii and one S. solfataricus. SpeciesFinder, TaxonomyFinder, and KmerFinder predicted the species of all three isolates correctly, while rMLST, which was only intended for characterization of Bacteria (15) predicted the M. smithii correctly, but was unable to make a prediction for the S. solfataricus.

The overlap in predictions of the four methods was examined and illustrated in Figure 2A. All four methods correctly identified 428 out of 695 isolates (62%), and all methods misidentified the same six isolates. Table 1 lists these six isolates. Since all four methods agreed on these predictions, the isolates are likely to be wrongly annotated. Alternatively, the annotations of the isolates in the training data that the predictions were based on are incorrect.

As seen in Figure 2A, isolate predictions agreed upon by several methods are more accurate that predictions unique to a particular method. However, the KmerFinder method made unique predictions for 36 isolates of which 20 were in concordance with the annotation.

Predictions for the most common species in the dataset were examined more closely and illustrated in Figure 3 and in Supplementary Figure 2-5. In general, the wrong predictions by SpeciesFinder (that is, the ones that were in disagreement with the NCBI annotation) were typically scattered, often consisting of a few wrong predictions of each type. The rMLST method was, on the other hand, more consistent in its incorrect predictions. As an example, the rMLST method wrongly annotated all 14 Bacillus anthracis isolates as Bacillus thuringiensis, all 8 Brucella abortus as Brucella suis, and all 6 Burkholderia mallei as Burkholderia pseudomallei. In general, all four methods had difficulties identifying species within the Bacillus genus, such as isolates annotated as B. thuringiensis, but predicted to be Bacillus cereus or vice versa. Another mistake common to all methods was Streptococcus miltis being predicted as Streptococcus oralis or Streptococcus pneumoniae. Also, none of

Table 1: Isolates of the NCBI drafts set for which all four methods predict the species to be different from what it is annotated as.

RefSeqID	Strain name	Annotated species	Predicted species
NZ_ACLX00000000	AH621 uid55161	Bacillus cereus	Bacillus weihenstephanensis
NZ_ACMD00000000	BDRD ST196 uid55169	Bacillus cereus	Bacillus weihenstephanensis
NZ_ABDQ00000000	C Eklund uid54841	Clostridium botulinum	Clostridium novyi
NZ_ABXZ00000000	FTG uid55313	Francisella novicida	Francisella tularensis
NZ_AHIE00000000	DC283 uid86627	Pantoea stewartii	Pantoea ananatis
NZ_AEPO00000000	ATCC 49296 uid61461	Streptococcus sanguinis	Streptococcus oralis

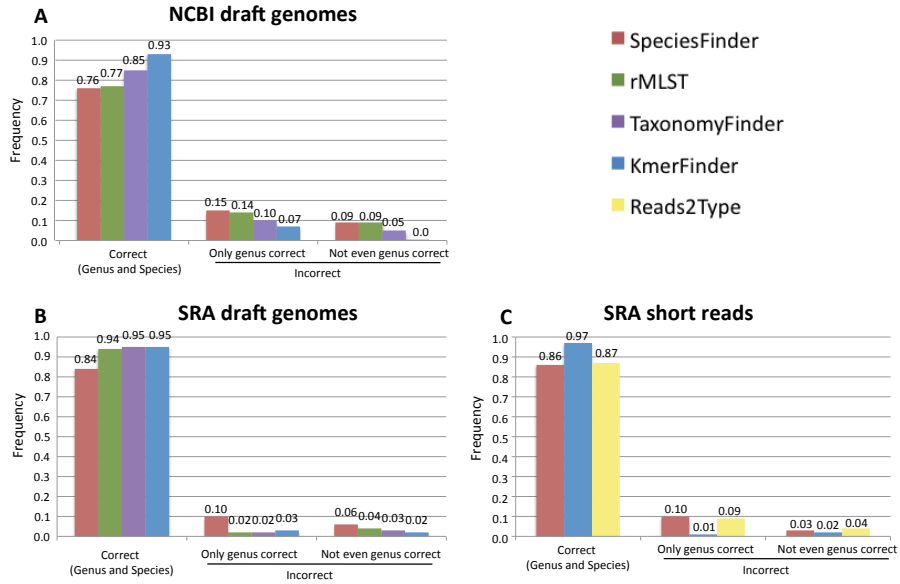


Figure 1: Performance of the five methods for species identification on A : NCB $_{drafts}$ B : SRA $_{drafts}$ C : SRA $_{reads}$. The rMLST and TaxonomyFinder methods only take draft or complete genomes as input, while Reads2Type only works for short reads. "Correct (genus and species)": Predicted genus and species are in accordance with the annotation. "Only genus correct": The predicted genus is in accordance with the annotation, but the species is not. "Not even genus correct": Neither predicted genus nor species is in accordance with the annotation.

the methods were able to correctly identify all annotated E. coli isolates, but identified at least some of them as Shigella spp. SpeciesFinder and TaxonomyFinder both had problems identifying the Borrelia burgorferi isolates, while SpeciesFinder and rMLST had problems distinguishing Yersinia pestis from Yersinia pseudotuberculosis . SpeciesFinder was the only method that had difficulties identifying Mycobacterium tuberculosis isolates, often predicting them to be Mycobacterium bovis.

Performances on SRA draft genomes

The SpeciesFinder, rMLST, TaxonomyFinder, and KmerFinder methods were next evaluated on the SRAdrafts set of 10,407 draft genomes covering 167 species. The performances on the draft genomes, for which the methods were able to make a prediction, are depicted in Figure 1B, while the overlap in predictions is illustrated in Figure 2B. Again, SpeciesFinder had the lowest performance with only 84% correct predictions. The rMLST, TaxonomyFinder, and KmerFinder methods had almost equal performances of 94%, 95%, and 95%, respectively. There was, however, a difference in the percentage of draft genomes for which each of the methods failed to make any prediction. SpeciesFinder and KmerFinder were the most robust methods, failing to make predictions for only 0.2% and 0.4% of the draft genomes, respectively. TaxonomyFinder was not able to make a prediction for 1.8% of the draft genomes, and rMLST not for 3.5%. That rMLST was the least robust method was at least partly due to our implementation of the method, where only hits with at least 95% identity and

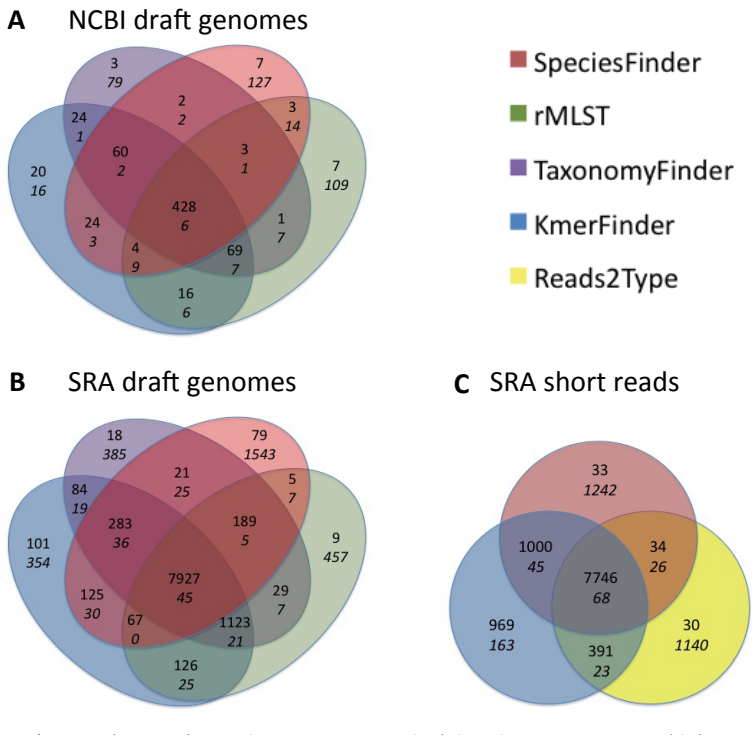


Figure 2: Overlap in predictions by the five methods for species identification. Numbers written in regular font indicate the number ofi solates for which the predicted species corresponds to the annotated species. Numbers written in italics indicate the number ofi solates for which the predicted and annotated species differ. A : The 16S, rMLST, KmerFinder and TaxonomyFinder methods evaluated on the NCBI $_{drafts}$ set. B : The 16S, rMLST, and KmerFinder methods evaluated on the SRA $_{drafts}$ set. C : The 16S, KmerFinde r, and Reads2Type methods evaluated on the SRA $_{reads}$ set.

95% coverage were considered a potential match. On the other hand, the N50 values for the draft genomes that SpeciesFinder and KmerFinder could not make a prediction for, were approximately half the size of the corresponding values for rMLST and TaxonomyFinder (data not shown), meaning that the quality of the draft genomes had to be higher for rMLST and TaxonomyFinder to be able to make a prediction. This is in accordance with these methods relying on the presence of many complete genes in the draft genomes.

Predictions for the most common species in the dataset are shown in Figure 4 and in Supplementary Figure 6-9. As seen previously when evaluating on the NCBI $_{drafts}$ set, the rMLST method was more consistent in its predictions for a given species than the other methods. For instance, rMLST predicted all 15 Mycobacterium bovis isolates to be M. tuberculosis. As also seen when evaluating on the NCBI $_{drafts}$ set, it is evident that all methods had difficulties distinguishing E. coli from species within the Shigella genus. Furthermore, species within the Brucella genus were often wrongly identified. In particular, it was only TaxonomyFinder that was able to correctly identify most Brucella abortus isolates. Some of the common problems that were obvious when evaluating on the NCBI $_{drafts}$ set, were not obvious when evaluating on the SRA $_{drafts}$ set, since the problematic species were too scarcely represented here. For instance, there are only five species from the Bacillus genus and only one S. mitis in SRA $_{drafts}$. The difference in species distribution between the NCBI $_{drafts}$ and SRA $_{drafts}$ set

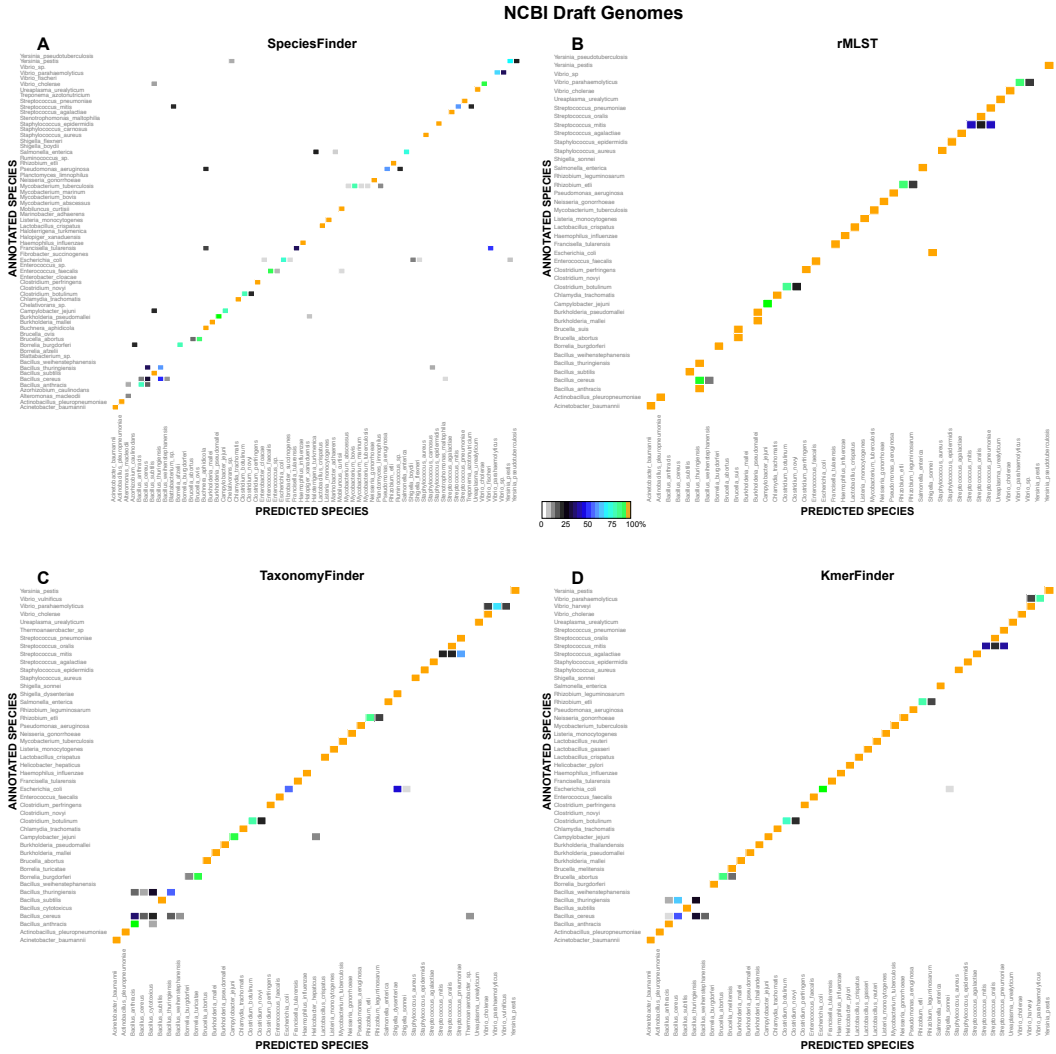


Figure 3: Predictions for the most common species of the NCBI $_{\text{drafts}}$ set. For each method, the results for a given species is only shown if the method made a prediction for five or more isolates annotated as this species (e.g., if there are five isolates annotated as species A in the dataset, but the method was not able to make a prediction for one of the isolates, the species is not shown), or two or more isolates are predicted as this species (e.g., there are no isolates annotated as species B in the dataset, but two isolates annotated as species C are predicted to be species B, then species B is shown). A : Predictions by SpeciesFinder. B : Predictions by rMLST. C : Predictions by TaxonomyFinder. D : Predictions by KmerFinder.

also explain why SpeciesFinder, TaxonomyFinder and rMLST all have increased performance on the SRA $_{drafts}$ set: While more than half of the isolates in the SRA $_{drafts}$ set belong to the Salmonella, Staphylococcus or Streptococcus genera, which none of the methods have particular problems identifying, these genera constitute less than 20% of NCBI $_{drafts}$. Conversely, the NCBI $_{drafts}$ set contains a high proportion of the problematic species E. coli (8.8%) and the genus Bacillus (10%). The corresponding proportions for SRA $_{drafts}$ are 3.5% E. coli and 0.05% isolates of the Bacillus genus. Furthermore, the NCBI $_{drafts}$ set is proportionally more diverse consisting of 149 species, while the almost 15 times larger SRA $_{drafts}$ set consists of only 168 different species.

Performances on short reads from SRA

Only three of the methods were able to perform species predictions directly on short reads, without first assembling the reads. These methods were SpeciesFinder, KmerFinder, and Reads2Type. Their performances on the SRA $_{reads}$ set of 10,407 sets of short reads representing 168 species are shown in Figure 1C.

Again, the SpeciesFinder method had the poorest performance with 86% of the isolates being correctly predicted. Reads2Type performed a bit better (87%), while KmerFinder achieved 97% correct.

Figure 2C illustrates the overlap in predictions between the three methods, while predictions for the most common species are shown in Supplementary Figure 10. In general, the results correspond to those observed for the SRAdrafts set.

Speed

The speed of the methods was evaluated on a subset of draft genomes and short reads as described in the Material and Methods. Since the actual speed experienced by the user will depend on a number off actors, for instance, the network bandwidth capacity of the client computer and the number ofj obs queued at the server, the relative speed of the different methods in comparison to each other is more relevant than the absolute speed.

Table 2: Speed of the tested methods.

Method	Speed on draft genomes	Speed on short reads
SpeciesFinder	00:13	3:14
Reads2Type	NA	1:20
rMLST	00:45	NA
TaxonomyFinder	11:33	NA
KmerFinder	00:09	03:10

DISCUSSION

In the present study we trained five different methods for prokaryotic species identification on a common dataset and evaluated their performances on three datasets of draft genomes

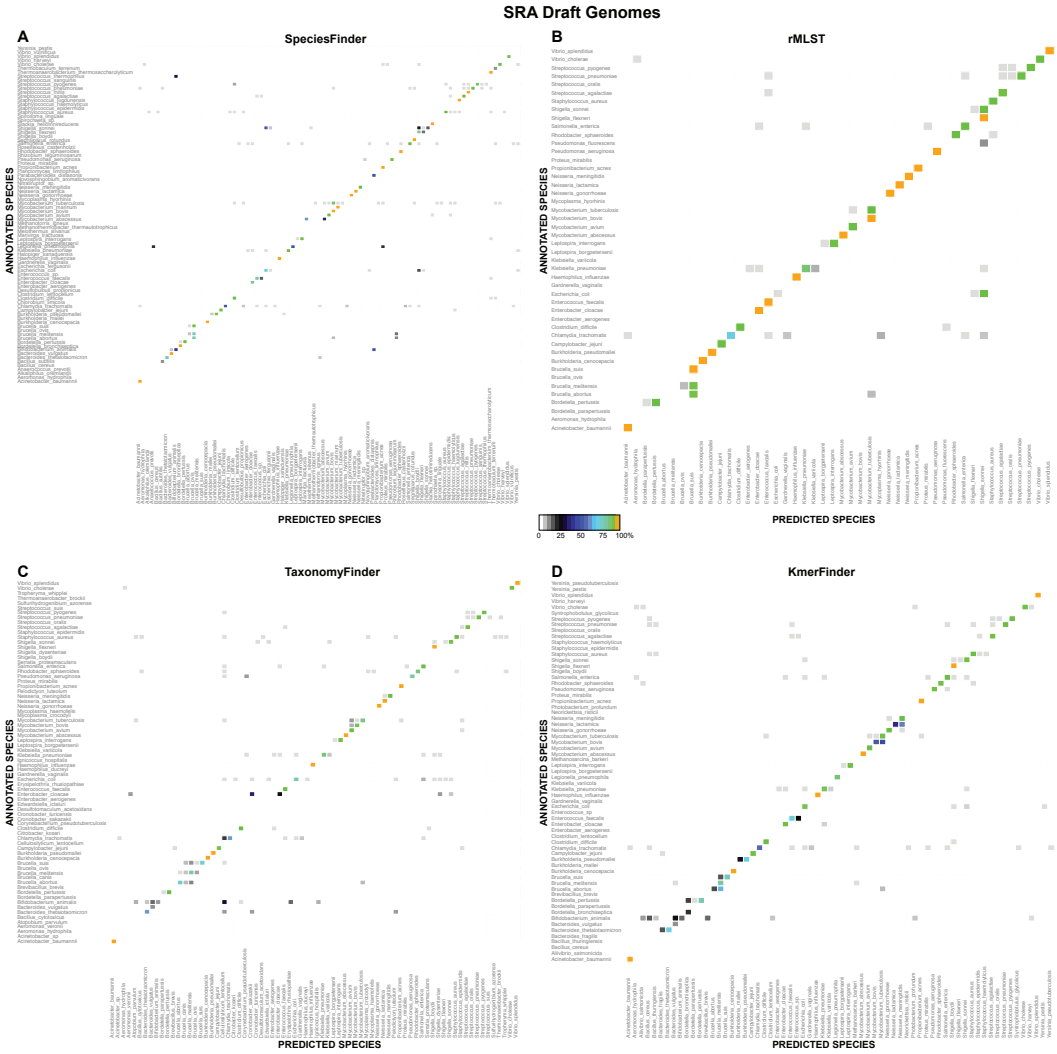


Figure 4: Predictions for the most common species in the SRA $_{\text{drafts}}$ dataset. For each method, the results for a given species is only shown if the method made a prediction for ten or more isolates annotated as this species, or two or more isolates are predicted as this species A: Predictions by SpeciesFinder. B: Predictions by rMLST. C: Predictions by TaxonomyFinder. D: Predictions by KmerFinder.

or short sequence reads.

The SpeciesFinder method is based on the 16S rRNA gene, which has served as the backbone of prokaryotic systematics since 1977 (1). Accordingly, sequencing of the 16S rRNA gene is a well-established method for identification of prokaryotes and has in all likelihood been used for annotating some of the isolates in the training and evaluation sets. In the light of this potential advantage of the SpeciesFinder method over the other methods, it is noteworthy that it had the lowest performance on all evaluation sets. Previous studies have, however, also pointed to the many limitations of the 16S rRNA gene for taxonomic purposes. Examples, which are also observed in this study, include its inadequacy for the delineation of species within the Borrelia burgdorferi sensu lato complex and the Mycobacterium tuberculosis complex (32). Similarly, in silico studies of the applicability of the 16S rRNA gene for the identification of medically important bacteria led to the authors concluding that although the method is useful for identification to the genus level, it is only able to identify 62% of anaerobic bacteria (33) and less than 30% of aerobic bacteria (34) confidently to the species level.

The performance of SpeciesFinder was surpassed only marginally by Reads2Type. This is not surprising, since the two methods are conceptionally very similar: SpeciesFinder utilizes the entire 16S rRNA gene of approximately 1,540 nucleotides, while for most species, Reads2Type looks for species-specific 50-mers in the same gene. In terms of its future usability, Reads2Type has, however, one advantage over the other methods: Like most of the other methods it is available as a web-server, but uniquely it does not require the read data to be uploaded to the server. Instead, a small 50-mer database is transferred to the user's computer and all computations performed here. As a result, bottleneck problems on the server are avoided and the data transfer is minimized, which may be particularly advantageous for users with limited Internet access.

While SpeciesFinder and Reads2Type only sample one locus, the rMLST method samples up to 53 loci – all ribosomal genes located to the chromosome of the bacteria. Evaluating on the dataset of SRA draft genomes, rMLST, TaxonomyFinder, and KmerFinder performed equally well. However, on the more diverse and difficult set of NCBI draft genomes, the rMLST method performed only marginally better than SpeciesFinder and significantly worse than TaxonomyFinder and KmerFinder. In particular, the rMLST method consistently made incorrect identifications of a number of closely related species, e.g., Y. pestis versus Y. pseudotuberculosis (35) and M. tuberculosis versus M. bovis (36). Also, rMLST consistently predicted the human pathogen B. anthracis to be B. thuringiensis . The later is used extensively as a biological pesticide and is generally not considered harmful for humans. B. anthracis and B. thuringiensis are both members of the B. cereus group and genetically very similar, with most of the disease and host specificity being attributable to their content of plasmids (37; 38). It has even been suggested that all members of the B. cereus group should be considered to be B. cereus and only subsequently be differentiated by their plasmids (39). Hence, in concordance with rMLST sampling only chromosomal, core genes, it is not surprising that the method fails to distinguish these isolates. A similar example is given by the rMLST method identifying all E. coli isolates as Shigella sonnei. Although Shigella spp. isolates have been rewarded their own genus, its separation from Escherichia spp. is mainly historical (40; 41; 42). To be sure, some of the mistakes commonly made by rMLST as well as the other methods highlight taxonomic taxa that are intrinsically difficult to distinguish due to a sub-optimal initial classification: Although Shigella spp. has for several years been considered a sub-strain of E. coli , the practical implications of renaming it is considered insurmountable.

The TaxonomyFinder method was the second most accurate method on the set of NCBI draft genomes and performed in the top for the SRA_{drafts} set. In contrary to the other methods it does not work directly on the nucleotide sequence of the isolates, but rather on

the proteome, utilizing functional protein domain profiles for the species prediction. It was the slowest of the tested methods, but in return for the extra time, the user is rewarded with an annotated genome.

The KmerFinder method performs its predictions on the basis of co-occurring k-mers, regardless of their location in the chromosome. It had the overall highest accuracy, works on complete or draft genomes as well as short reads, was found to be very robust as well as fast. Furthermore, the KmerFinder method holds promise for future improvements, as the implementation used for this study was very simple: Only the raw number of co-occuring k-mers between the query and reference genome was considered, although a parallel analysis indicates that the performance could be improved even further if more sophisticated measures were used, also taking into account the total number of k-mers in the query and reference genome.

It has previously been noted that some of the isolates present in public databases, and hence used in this study, are wrongly annotated (16; 43; 44). Based on the current study, it is likely that at least the six isolates from the NCBI $_{\text{drafts}}$ set that all methods identified as something different than the annotated species, are wrongly annotated. In agreement with this, one of the isolates has indeed been re-annotated, since we initially downloaded the data. Of the remaining five isolates, two *B. cerues* isolates were found to be most closely related to the *B. weihenstephanensis* strain KBAB4 of the common training set. This strain is the single representative of the species in the public database and not the type strain. Hence there is no guarantee that the sequenced strain represents the named taxon (45). The same is the case for the *C. botulinum* strain C Eklund, which is predicted to be a *Clostridium novyi* based on its close resemblance to *C. novyi* strain NT of the training set. *Clostridium novyi* strain NT is the only representative of this species in the database and not the type strain.

While some taxonomists consider the goal of bacterial taxonomy to mirror the order of nature and describe the evolutionary order back to the origin ofl ife (5; 46), a more pragmatic and applied view is likely to be advantageous for epidemiological purposes, where most outbreaks last less than six months. The number of prokaryotic genomes in public databases is currently sufficiently high to substitute theoretical views of which loci to sample for optimal species identification by actual testing how different approaches perform. One locus (the 16S rRNA gene) was initially used for sequenced-based examination of relationships between bacteria, and when the approach was found to have limitations, more loci were added in MLST and MLSA (47; 48). The addition of still more loci has been suggested for improving MLSA even further (32; 15). This study suggests that an optimal approach should not be limited to a finite number of genes, but rather look at the entire genome.

CONCLUSION

The 16S rRNA gene has served prokaryotic taxonomy well for more than 30 years, but the emergence of second- and third generation sequencing technologies enables the use of WGS data with the potential of higher resolution and more phylogenetically accurate classifications. Methods that sample the entire genome, not just core genes located to the chromosome, seems particularly well suited for taking up the baton.

ACKNOWLEDGEMENTS

This work was supported by the Center for Genomic Epidemiology at the Technical University of Denmark and funded by grant 09-067103/DSF from the Danish Council for Strategic Research.
We are grateful to John Damm Sørensen for excellent technical assistance. We are grateful to Keith Jolley, Department of Zoology, University of Oxford, UK for providing us with the rMLST genes for the genomes in the training data.

References

[1] G. E. FOX, K. R. PECHMAN, and C. R. WOESE, "Comparative Cataloging of 16S Ribosomal Ribonucleic Acid: Molecular Approach to Procaryotic Systematics," International Journal of Systematic Bacteriology , vol. 27, pp. 44–57, Jan. 1977.

[2] T. Z. DeSantis, P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen, "Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB.," Applied and environmental microbiology, vol. 72, pp. 5069–72, July 2006.

[3] W. Ludwig, O. Strunk, R. Westram, L. Richter, H. Meier, Yadhukumar, A. Buchner, T. Lai, S. Steppi, G. Jobb, W. Förster, I. Brettske, S. Gerber, A. W. Ginhart, O. Gross, S. Grumann, S. Hermann, R. Jost, A. König, T. Liss, R. Lüssmann, M. May, B. Nonhoff, B. Reichel, R. Strehlow, A. Stamatakis, N. Stuckmann, A. Vilbig, M. Lenke, T. Ludwig, A. Bode, and K.-H. Schleifer, "ARB: a software environment for sequence data.," Nucleic acids research , vol. 32, pp. 1363–71, Jan. 2004.

[4] E. Pruesse, C. Quast, K. Knittel, B. M. Fuchs, W. Ludwig, J. Peplies, and F. O. Glöckner, "SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB.," Nucleic acids research , vol. 35, pp. 7188–96, Jan. 2007.

[5] P. Kämpfer, "Systematics of prokaryotes: the state of the art.," Antonie van Leeuwenhoek, vol. 101, pp. 3–11, Jan. 2012.

[6] B. J. Tindall, R. Rosselló-Móra, H.-J. Busse, W. Ludwig, and P. Kämpfer, "Notes on the characterization of prokaryote strains for taxonomic purposes.," International journal of systematic and evolutionary microbiology, vol. 60, pp. 249–66, Jan. 2010.

[7] B. J. Tindall, S. Schneider, A. Lapidus, A. Copeland, T. Glavina Del Rio, M. Nolan, S. Lucas, F. Chen, H. Tice, J.-F. Cheng, E. Saunders, D. Bruce, L. Goodwin, S. Pitluck, N. Mikhailova, A. Pati, N. Ivanova, K. Mavrommatis, A. Chen, K. Palaniappan, P. Chain, M. Land, L. Hauser, Y.-J. Chang, C. D. Jeffries, T. Brettin, C. Han, M. Rohde, M. Göker, J. Bristow, J. A. Eisen, V. Markowitz, P. Hugenholtz, H.-P. Klenk, N. C. Kyrpides, and J. C. Detter, "Complete genome sequence of Halomicrobium mukohataei type strain (arg-2).," Standards in genomic sciences, vol. 1, pp. 270–7, Jan. 2009.

[8] M. Walcher, R. Skvoretz, M. Montgomery-Fullerton, V. Jonas, and S. Brentano, "Description of an Unusual Neisseria meningitidis Isolate Containing and Expressing Neisseria gonorrhoeae-Specific 16S rRNA Gene Sequences.," Journal of clinical microbiology , vol. 51, pp. 3199–206, Oct. 2013.

[9] H.-P. Klenk and M. Göker, "En route to a genome-based classification of Archaea and Bacteria?," Systematic and applied microbiology, vol. 33, pp. 175–82, June 2010.

[10] B. Snel, P. Bork, and M. A. Huynen, "Genome phylogeny based on gene content.," Nature genetics, vol. 21, pp. 108–10, Jan. 1999.

[11] C. H. House and S. T. Fitz-Gibbon, "Using homolog groups to create a whole-genomic tree off ree-living organisms: an update.," Journal of molecular evolution , vol. 54, pp. 539–47, Apr. 2002.

[12] S. Yang, R. F. Doolittle, and P. E. Bourne, "Phylogeny determined by protein domain content.," Proceedings of the National Academy of Sciences of the United States of America , vol. 102, pp. 373–8, Jan. 2005.

[13] O. Lukjancenko, M. C. Thomsen, M. V. Larsen, and D. W. Ussery, "PanFunPro: PAN-genome analysis based on FUNctional PROfiles," submitted to F1000Research , 2013.

[14] F. D. Ciccarelli, T. Doerks, C. von Mering, C. J. Creevey, B. Snel, and P. Bork, "Toward automatic reconstruction of a highly resolved tree ofl ife.," Science (New York, N.Y.) , vol. 311, pp. 1283–7, Mar. 2006.

[15] K. A. Jolley, C. M. Bliss, J. S. Bennett, H. B. Bratcher, C. Brehony, F. M. Colles, H. Wimalarathna, O. B. Harrison, S. K. Sheppard, A. J. Cody, and M. C. J. Maiden, "Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain.," Microbiology (Reading, England) , vol. 158, pp. 1005–15, Apr. 2012.

[16] J. S. Bennett, K. A. Jolley, S. G. Earle, C. Corton, S. D. Bentley, J. Parkhill, and M. C. J. Maiden, "A genomic approach to bacterial taxonomy: an examination and proposed reclassification of species within the genus Neisseria.," Microbiology (Reading, England) , vol. 158, pp. 1570–80, June 2012.

[17] A. J. Cody, N. D. McCarthy, M. Jansen van Rensburg, T. Isinkaye, S. D. Bentley, J. Parkhill, K. E. Dingle, I. C. J. W. Bowler, K. A. Jolley, and M. C. J. Maiden, "Real-time genomic epidemiological evaluation of human Campylobacter isolates by use of whole-genome multilocus sequence typing.," Journal of clinical microbiology , vol. 51, pp. 2526–34, Aug. 2013.

[18] Y. Kodama, M. Shumway, and R. Leinonen, "The Sequence Read Archive: explosive growth of sequencing data.," Nucleic acids research , vol. 40, pp. D54–6, Jan. 2012.

[19] D. R. Zerbino and E. Birney, "Velvet: algorithms for de novo short read assembly using de Bruijn graphs.," Genome research, vol. 18, pp. 821–9, May 2008.

[20] M. V. Larsen, S. Cosentino, S. Rasmussen, C. Friis, H. Hasman, R. L. Marvig, L. Jelsbak, T. Sicheritz-Pontén, D. W. Ussery, F. M. Aarestrup, and O. Lund, "Multilocus sequence typing of total-genome-sequenced bacteria.," Journal of clinical microbiology , vol. 50, pp. 1355–61, Apr. 2012.

[21] J. R. Miller, S. Koren, and G. Sutton, "Assembly algorithms for next-generation sequencing data.," Genomics, vol. 95, pp. 315–27, June 2010.

[22] K. Lagesen, P. Hallin, E. A. Rø dland, H.-H. Staerfeldt, T. r. Rognes, and D. W. Ussery, "RNAmmer: consistent and rapid annotation of ribosomal RNA genes.," Nucleic acids research, vol. 35, pp. 3100–8, Jan. 2007.

[23] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform.," Bioinformatics (Oxford, England) , vol. 25, pp. 1754–60, July 2009.

[24] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev, "Full-length transcriptome assembly from RNA-Seq data without a reference genome.," Nature biotechnology, vol. 29, pp. 644–52, July 2011.

[25] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.," Nucleic acids research , vol. 25, pp. 3389–402, Sept. 1997.

[26] W. J. Kent, "BLAT–the BLAST-like alignment tool.," Genome research, vol. 12, pp. 656–64, Apr. 2002.

[27] M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, A. Bateman, and R. D. Finn, "The Pfam protein families database.," Nucleic acids research , vol. 40, pp. D290–301, Jan. 2012.

[28] D. H. Haft, J. D. Selengut, and O. White, "The TIGRFAMs database of protein families.," Nucleic acids research , vol. 31, pp. 371–3, Jan. 2003.

[29] J. Gough, K. Karplus, R. Hughey, and C. Chothia, "Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure.," Journal of molecular biology , vol. 313, pp. 903–19, Nov. 2001.

[30] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.," Bioinformatics (Oxford, England) , vol. 22, pp. 1658–9, July 2006.

[31] D. Hyatt, G.-L. Chen, P. F. Locascio, M. L. Land, F. W. Larimer, and L. J. Hauser, "Prodigal: prokaryotic gene recognition and translation initiation site identification.," BMC bioinformatics , vol. 11, p. 119, Jan. 2010.

[32] L. A. Almeida and R. Araujo, "Highlights on molecular identification of closely related species.," Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases , vol. 13, pp. 67–75, Jan. 2013.

[33] P. C. Y. Woo, L. M. W. Chung, J. L. L. Teng, H. Tse, S. S. Y. Pang, V. Y. T. Lau, V. W. K. Wong, K.-l. Kam, S. K. P. Lau, and K.-Y. Yuen, "In silico analysis of 16S ribosomal RNA gene sequencing-based methods for identification of medically important anaerobic bacteria.," Journal of clinical pathology , vol. 60, pp. 576–9, May 2007.

[34] J. L. L. Teng, M.-Y. Yeung, G. Yue, R. K. H. Au-Yeung, E. Y. H. Yeung, A. M. Y. Fung, H. Tse, K.-Y. Yuen, S. K. P. Lau, and P. C. Y. Woo, "In silico analysis of 16S rRNA gene sequencing based methods for identification of medically important aerobic Gram-negative bacteria.," Journal of medical microbiology , vol. 60, pp. 1281–6, Sept. 2011.

[35] M. Achtman, K. Zurth, G. Morelli, G. Torrea, A. Guiyoule, and E. Carniel, "Yersinia pestis, the cause of plague, is a recently emerged clone of Yersinia pseudotuberculosis.," Proceedings of the National Academy of Sciences of the United States of America , vol. 96, pp. 14043–8, Nov. 1999.

[36] S. Sreevatsan, X. Pan, K. E. Stockbauer, N. D. Connell, B. N. Kreiswirth, T. S. Whittam, and J. M. Musser, "Restricted structural gene polymorphism in the Mycobacterium tuberculosis complex indicates evolutionarily recent global dissemination.," Proceedings of the National Academy of Sciences of the United States of America , vol. 94, pp. 9869–74, Sept. 1997.

[37] G. Jiménez, M. Urdiain, A. Cifuentes, A. López-López, A. R. Blanch, J. Tamames, P. Kämpfer, A.-B. Kolstø, D. Ramón, J. F. Martínez, F. M. Codoñer, and R. Rosselló-Móra, "Description of Bacillus toyonensis sp. nov., a novel species of the Bacillus cereus group, and pairwise genome comparisons of the species of the group by means of ANI calculations.," Systematic and applied microbiology, vol. 36, pp. 383–91, Sept. 2013.

[38] D. A. Rasko, M. R. Altherr, C. S. Han, and J. Ravel, "Genomics of the Bacillus cereus group of organisms.," FEMS microbiology reviews , vol. 29, pp. 303–29, Apr. 2005.

[39] E. Helgason, O. A. Okstad, D. A. Caugant, H. A. Johansen, A. Fouet, M. Mock, I. Hegna, and A. B. Kolstø, "Bacillus anthracis, Bacillus cereus, and Bacillus thuringiensis–one species on the basis of genetic evidence.," Applied and environmental microbiology, vol. 66, pp. 2627–30, June 2000.

[40] D. K. Karaolis, R. Lan, and P. R. Reeves, "Sequence variation in Shigella sonnei (Sonnei), a pathogenic clone of Escherichia coli, over four continents and 41 years.," Journal of clinical microbiology , vol. 32, pp. 796–802, Mar. 1994.

[41] R. Lan and P. R. Reeves, "Escherichia coli in disguise: molecular origins of Shigella.," Microbes and infection / Institut Pasteur , vol. 4, pp. 1125–32, Sept. 2002.

[42] O. Lukjancenko, T. M. Wassenaar, and D. W. Ussery, "Comparison of 61 sequenced Escherichia coli genomes.," Microbial ecology, vol. 60, pp. 708–20, Nov. 2010.

[43] J. Goris, K. T. Konstantinidis, J. A. Klappenbach, T. Coenye, P. Vandamme, and J. M. Tiedje, "DNA-DNA hybridization values and their relationship to whole-genome sequence similarities.," International journal of systematic and evolutionary microbiology , vol. 57, pp. 81–91, Jan. 2007.

[44] P. Yarza, M. Richter, J. Peplies, J. Euzeby, R. Amann, K.-H. Schleifer, W. Ludwig, F. O. Glöckner, and R. Rosselló-Móra, "The All-Species Living Tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains.," Systematic and applied microbiology, vol. 31, pp. 241–50, Sept. 2008.

[45] M. Richter and R. Rosselló-Móra, "Shifting the genomic gold standard for the prokaryotic species definition.," Proceedings of the National Academy of Sciences of the United States of America , vol. 106, pp. 19126–31, Nov. 2009.

[46] P. Kämpfer and S. P. Glaeser, "Prokaryotic taxonomy in the sequencing era–the polyphasic approach revisited.," Environmental microbiology , vol. 14, pp. 291–317, Feb. 2012.

[47] D. Gevers, F. M. Cohan, J. G. Lawrence, B. G. Spratt, T. Coenye, E. J. Feil, E. Stackebrandt, Y. Van de Peer, P. Vandamme, F. L. Thompson, and J. Swings, "Opinion: Re-evaluating prokaryotic species.," Nature reviews. Microbiology , vol. 3, pp. 733–9, Sept. 2005.

[48] M. C. Maiden, J. A. Bygraves, E. Feil, G. Morelli, J. E. Russell, R. Urwin, Q. Zhang, J. Zhou, K. Zurth, D. A. Caugant, I. M. Feavers, M. Achtman, and B. G. Spratt, "Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms.," Proceedings of the National Academy of Sciences of the United States of America , vol. 95, pp. 3140–5, Mar. 1998.

4 BACTERIAL ANTIBIOTIC RESISTANCE

Antibiotics are biochemical agents that inhibit the growth or kill bacteria. The possible existence of antimicrobial drugs was first proposed in 1877 by Louis Pasteur and Robert Koch. In 1895, Vincenzo Tiberio, physician of the University of Naples discovered that a mold (*Penicillium*) had an antibacterial action, which was followed in 1923 by the discovery of Penicillin, by Alexander Fleming. The first commercially available antibiotic (*prontosil*) was discovered by a team of researchers led by Gerhard Domagk in 1932 at the Bayer Laboratories in Germany. This event marked the start of the antibiotic era, with has a period rich with discoveries of many synthetic and natural antimicrobial drugs with its peak around 70s and a dramatic decrease continuing in present days. When antimicrobial drugs were discovered many scientists and clinicians thought that all infections could be cured, but they did not consider the ability of bacteria to quickly evolve and become resistant to antimicrobial drugs.

Nowadays AMR is considered one of the biggest health challenges of 21^{th} century, with the return of diseases like gonorrhoea (from *Neisseria gonorrhoeae*), pneumonia (from *Klebsiella pneumoniae*) and meningitis (from *Neisseria meningitis*) becoming a threat again mainly because the bacteria causing it are becoming resistant to most of the antimicrobial drugs available in the market. Among the causes behind the increase of antimicrobial resistance [74] there is a misuse of antibiotics from people [48] (Figure 4.1), due also to the relatively easy access to these drugs and inaccurate prescriptions from medics, and also the, sometimes poorly controlled, use of antibiotics in agriculture and livestock's food [45] (Figure 5). Studying bacterial antimicrobial resistance is very important in order to understand the mechanisms behind the resistance of AMR strains and find different treatments for the diseases caused by these bacteria. In the scenario of infections causes by novel or mutated strains, to quickly understand what antibiotic the bacteria is resistant to could help medics in choosing the proper drug to cure the infected patient, increasing the chances of saving the patient's life and, at the same time, decreasing the possibility of potential outbreaks to spread. Assessing bacterial antibiotic resistance is still expensive and time consuming, and in the manuscript accompanying this chapter [76], we propose ResFinder, a free to use web-based tool (*http://cge.cbs.dtu.dk/services/ResFinder/*) for the fast identification of acquired antimicrobial resistant genes from WGS data.

ResFinder is at present the most used service among those offered by CGE, with about 1000 (and growing) jobs served every month with users from more than 60 countries. The web-service is proving to be very useful not only for researchers in developed countries with high-occurrence of AMR-related infections, but also for scientists and clinicians in developing countries,

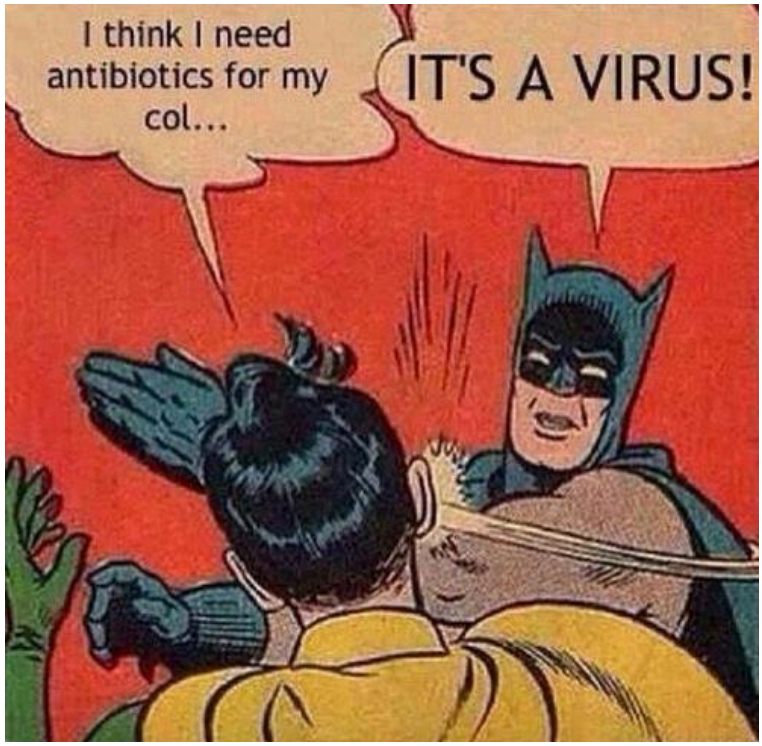


Figure 4.1: the use of antibiotics for curing diseases in which bacteria are not involved is one of the causes of the increase of microbial resistance in bacteria [59, 60]

where the costs of traditional antibiotic resistance assessment are a limiting factor in this research.

Manuscript IV

Identification of acquired antimicrobial resistance genes

J Antimicrob Chemother 2012; **67**: 2640–2644
doi:10.1093/jac/dks261 Advance Access publication 10 July 2012

Journal of Antimicrobial Chemotherapy

Identification of acquired antimicrobial resistance genes

Ea Zankari[1,2]*, Henrik Hasman[1], Salvatore Cosentino[2], Martin Vestergaard[1], Simon Rasmussen[2], Ole Lund[2], Frank M. Aarestrup[1] and Mette Voldby Larsen[2]

[1]*National Food Institute, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark;* [2]*Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark*

*Corresponding author. Tel: +45-35887183; E-mail: east@food.dtu.dk

Received 13 March 2012; returned 26 April 2012; revised 8 June 2012; accepted 13 June 2012

Objectives: Identification of antimicrobial resistance genes is important for understanding the underlying mechanisms and the epidemiology of antimicrobial resistance. As the costs of whole-genome sequencing (WGS) continue to decline, it becomes increasingly available in routine diagnostic laboratories and is anticipated to substitute traditional methods for resistance gene identification. Thus, the current challenge is to extract the relevant information from the large amount of generated data.

Methods: We developed a web-based method, ResFinder that uses BLAST for identification of acquired antimicrobial resistance genes in whole-genome data. As input, the method can use both pre-assembled, complete or partial genomes, and short sequence reads from four different sequencing platforms. The method was evaluated on 1862 GenBank files containing 1411 different resistance genes, as well as on 23 *de-novo*-sequenced isolates.

Results: When testing the 1862 GenBank files, the method identified the resistance genes with an ID=100% (100% identity) to the genes in ResFinder. Agreement between *in silico* predictions and phenotypic testing was found when the method was further tested on 23 isolates of five different bacterial species, with available phenotypes. Furthermore, ResFinder was evaluated on WGS chromosomes and plasmids of 30 isolates. Seven of these isolates were annotated to have antimicrobial resistance, and in all cases, annotations were compatible with the ResFinder results.

Conclusions: A web server providing a convenient way of identifying acquired antimicrobial resistance genes in completely sequenced isolates was created. ResFinder can be accessed at www.genomicepidemiology.org. ResFinder will continuously be updated as new resistance genes are identified.

Keywords: antibiotic resistance, genotype, ResFinder, resistance gene identification

Introduction

The introduction of antimicrobial agents for treatment of infectious diseases is one of the most important achievements of the 20th century. However, soon after their introduction, isolates with acquired resistance emerged and this pattern has followed the introduction of each new antimicrobial agent.

A large number of different genes can be responsible for antimicrobial resistance. Identification of these genes is important to understand resistance epidemiology, for verification of non-susceptible phenotypes and for identification of resistant strains, when genes are weakly expressed *in vitro*. Detection of resistance genes has typically been performed using PCR[1] or microarrays.[2] However, in several cases, it is necessary to perform supplementary sequencing of the amplified PCR products.[3] As a result, it is expensive and time-consuming to perform a complete identification of resistance genes present in a strain collection.

The cost of DNA sequencing has steadily gone down, by roughly 10-fold every five years. As a consequence, DNA sequencing is becoming increasingly accessible for routine use and was recently utilized for complete characterization of antimicrobial resistance and virulence gene content during the safety evaluation of 28 strains intended for use in human nutrition.[4] The challenge is, however, to extract the relevant information from the large amount of data that is generated by these techniques.

The Center for Genomic Epidemiology (www.genomicepidemiology.org) aims at providing the bioinformatic and scientific foundation for processing and handling whole-genome sequencing (WGS) information in a standardized way useful for outbreak investigation, source tracking, diagnostics and epidemiological surveillance. The services are publically available through web servers specifically designed to be user-friendly—and also for investigators with limited bioinformatics experience.

We here present ResFinder, a web server that uses WGS data for identifying acquired antimicrobial resistance genes in bacteria.

Methods

Databases

Data on acquired resistance genes was collected from databases (http://faculty.washington.edu/marilynr/, http://ardb.cbcb.umd.edu/ and http://www.lahey.org/Studies/) and published papers including reviews.[5,6] All sequences were collected from the NCBI nucleotide database (http://www.ncbi.nlm.nih.gov/nuccore/) and used to build the ResFinder database. To our knowledge, we have created the largest collection of acquired antimicrobial resistance genes (see Table S1, available as Supplementary data at *JAC* Online).

Identifying resistance genes in completely sequenced bacteria

Draft assembly of short sequence reads was done as previously described.[7] All genes from the ResFinder database were BLASTed against the assembled genome, and the best-matching genes were given as output. For a gene to be reported, it has to cover at least 2/5 of the length of the resistance gene in the database. The best-matching genes were identified as previously[7]. It is possible to select a % identity (ID) threshold (the percentage of nucleotides that are identical between the best-matching resistance gene in the database and the corresponding sequence in the genome). The default ID is 100%.

Evaluation of method

Verification of the databases was made by testing ResFinder with the 1862 GenBank files from which the genes were collected, to verify that the method would find all genes with ID=100%.

Short sequence reads from 23 isolates of five different species, *Escherichia coli*, *Klebsiella pneumoniae*, *Salmonella enterica*, *Staphylococcus aureus* and *Vibrio cholerae*, were also submitted to ResFinder. All 23 isolates had been sequenced on the Illumina platform using paired-end reads. A ResFinder threshold of ID=98.00% was selected, as previous tests of ResFinder had shown that a threshold lower than this gives too much noise (e.g. fragments of genes). Phenotypic antimicrobial susceptibility testing was determined as MIC determinations, as previously described.[8]

With '(chromosome and plasmid)(multi-drug or antimicrobial or antibiotic)(resistant or resistance) pathogen' as search criteria, one isolate from each species with completely sequenced and assembled, and chromosome and plasmid data were collected from the NCBI Genomes database (http://www.ncbi.nlm.nih.gov/genome). This resulted in 30 isolates, from 30 different species, containing 85 chromosome/plasmid sequences. All sequences were run through all databases in ResFinder with a selected threshold of ID=98.00%.

Results

Using ResFinder

Short sequence reads can be assembled to draft genomes by the server. It is also possible to input a complete or partial, pre-assembled genome. ResFinder gives the option to run the input against one or several antimicrobial classes simultaneously, and it uses BLAST to identify the acquired resistance genes. It is possible to search for genes with specified similarity from 80%–100% identity, and the best-matching genes are given as output. An example of the output format is shown and explained at www.cbs.dtu.dk/services/ResFinder/output.php.

Evaluation of method

In all cases, ResFinder identified the acquired resistance genes in the 1862 GenBank files from which the databases were created, with an ID=100%.

Table 1 shows antimicrobial genes found by ResFinder, the predicted resistance profile and the phenotypic antimicrobial susceptibility test results for five bacterial isolates covering five different species. Tests for all 23 bacterial isolates covering the five different species can be seen in Table S2 (available as Supplementary data at *JAC* Online). Almost complete agreement between *in silico* predictions and phenotypic testing was found. The exceptions were two *S. aureus* isolates that contained the *mecA* gene but were phenotypically susceptible to penicillins, and two *S. aureus* isolates, one resistant to spectinomycin and the other to tiamulin, neither of which was found to contain genes matching these phenotypes. The *catB3* gene was found in all four *K. pneumoniae* isolates with an ID=100%, but not in full length, consistent with all four testing phenotypically susceptible to chloramphenicol. One *V. cholera* isolate contained part of *floR* and tested phenotypically susceptible to florfenicol.

Acquired antimicrobial resistance genes were found in 10 of the 30 strains from the NCBI genomes database (Table 2). For all except two isolates this coincided with the ResFinder results. *K. pneumoniae* KCTC 2242 was annotated to contain bla_{TEM}, whereas ResFinder detected bla_{SHV}. *Nocardia farcinica* IFM 10152 was annotated to contain a β-lactam gene as well as *aph(2″)*, *aph(3′)* and *aph(6)*, but ResFinder detected only the bla_{FAR-1} gene. These genes were further examined with BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi), which demonstrated that the genes detected by ResFinder were correct.

Discussion

Since their original development by Alexander Fleming, phenotypic disc diffusion and MIC determinations have been the gold standard for antimicrobial susceptibility testing. These methods have the great advantage of determining the 'true' *in vitro* relationship between the antimicrobial agent and the strain tested, and will detect any new emerging resistance mechanisms.

Genotypic testing of suspected resistant isolates is often performed to verify phenotypic observations and for epidemiological purposes. The most widely used approach has been to perform PCR to detect the presence of selected genes. In many cases only a single or a few genes mediating resistance are tested, and such studies will often miss the simultaneous presence of multiple genes encoding the same resistance.

WGS has the great benefit that it potentially provides complete information, and thus new experiments do not have to be performed to search for the presence of novel genes—the analysis can simply be rerun. One major obstacle is the lack of available bioinformatics tools allowing simple and standardized analysis of the large amounts of data generated by WGS.

We have developed, implemented and evaluated ResFinder, a method to detect the presence of 1862 different resistance genes from 12 different antimicrobial classes in WGS data (www.genomicepidemiology.org). The current version only

Table 1. ResFinder results for isolates of five different species compared with antimicrobial susceptibility data

Species	Isolate	ResFinder profile	Predicted phenotype	Detected phenotype
E. coli	Ødemsyge-186	*tet(A)*	TET	TET
K. pneumoniae	Kleb-6-1-264y	*aac(3)-IIa*[a]	GEN	GEN
		strA, *strB*	STR	STR
		$bla_{CTX\text{-}M\text{-}15}$	XNL, CTX, AMP	XNL, CTX, AMP
		$bla_{TEM\text{-}1}$	AMP	AMP
		$bla_{OXA\text{-}30}$	AMP, AMC	AMP, AMC
		$bla_{SHV\text{-}28}$	XNL, CTX, AMP	XNL, CTX, AMP
		aac(6′)Ib-cr	CIP	CIP[f]
		catB3[b]	CHL	—
		sul2	SMX	SMX
		tet(A)	TET	TET
		dfrA14[a]	TMP	TMP
		—	—	NAL[f]
S. enterica	Styph-0210H31581	*aac(6′)-Iaa*	[c]	—
		aadA2	SPT, STR	SPT, STR
		$bla_{CARB\text{-}2}$	AMP	AMP
		floR[a]	FFN, CHL	FFN, CHL
		sul1[b]	SMX	SMX
		tet(G)	TET	TET
		—	—	CIP[f], NAL[f]
S. aureus	2007-70-91-4	*aac(3)-Ik*[a]	[d]	—
		mecA	FOX	—
		blaZ	PEN	PEN
		tet(K), *tet(38)*[a], *tet(M)*[a]	TET	TET
		dfrG	TMP	TMP
		fusA[a]	FUS	—
V. cholerae	Vchole-002	*strA*, *strB*	STR	STR
		catB9	CHL[e]	—
		sul2	SMX	SMX
		dfrA1, *dfrA31*	TMP	TMP
		—	—	CIP[f], NAL[f], CST[f]

AMC, amoxicillin/clavulanate (2:1); AMP, ampicillin; CHL, chloramphenicol; CST, colistin; CTX, cefotaxime; FOX, cefoxitin; FFN, florfenicol; FUS, fusidic acid; GEN, gentamicin; PEN, penicillin; SMX, sulfamethoxazole; SPT, spectinomycin; STR, streptomycin; TET, tetracycline; TMP, trimethoprim; XNL, ceftiofur.

[a]The gene is found with an ID < 100%.

[b]The found gene is shorter than the resistance gene.

[c]Resistance to antimicrobials that were not included in the phenotypic antimicrobial susceptibility tests.

[d]Phenotype not known.

[e]Phenotypically silent in native position (19).

[f]Antimicrobial drug associated with chromosomal mutations.

covers horizontally acquired resistance genes and not resistance mediated by mutations, e.g. in housekeeping genes. ResFinder can also be used to ignore known acquired resistance genes in a search for new resistance genes.

ResFinder successfully identified all the genes from which the database was built, and correctly identified all genes present in 30 isolates of whole-genome data collected from the NCBI genomes database (http://www.ncbi.nlm.nih.gov/genome). Furthermore, phenotypic antimicrobial susceptibility tests of 23 isolates from five different species were compared with the results from ResFinder. With a few exceptions, complete agreement between predicted and observed phenotypes was found. All the *V. cholerae* isolates contained the *catB9* gene, which has previously been shown to be phenotypically silent in its native position,[9] consistent with all isolates testing phenotypically susceptible. The five *S. aureus* isolates examined in this study were from a collection of methicillin-resistant *S. aureus* (MRSA).[10] Phenotypic detection of *mecA*-harbouring isolates can be difficult, indicating the superiority of WGS compared with phenotypic testing. Two of the *S. aureus* isolates, 9B and PR11_08, showed phenotypic resistance to spectinomycin and tiamulin, respectively, but without containing any matching resistance genes. Interestingly, we found two extended-spectrum β-lactamase (ESBL)-related genes ($bla_{CTX\text{-}M\text{-}15}$ and $bla_{SHV\text{-}28}$) in

Table 2. ResFinder results for completely sequenced and assembled chromosome and plasmid data from 30 different species

Strain	Annotated resistance	Chromosome	Plasmid
Edwardsiella tarda EIB202	*tet(A)*, *tet(R)*, *strA*, *strB*, *sul2*	no genes found	*strA* 100% *strB* 100% *catA3* 99.84% *sul2* 100%; *tet(A)* 99.92%
Enterobacter cloacae subsp. *cloacae* ATCC 13047	—	*sul2* 100%	no genes found
Enterococcus faecalis	*tet(M)*	*tet(M)* 100%	no genes found
Fusobacterium nucleatum subsp. *polymorphum* ATCC 10953	—	*tet(K)* 100%	no genes found
Klebsiella pneumoniae KCTC 2242	β-lactam (bla_{TEM})	bla_{SHV-99} 99.88%	no genes found
Nocardia farcinica IFM 10152	β-lactam, *aph(2″)*, *aph(3′)*, *aph(6)*	bla_{FAR-1} 98.77%	no genes found
Ochrobactrum anthropi ATCC 49188	—	bla_{OCH-3} 99.91%	no genes found
Ralstonia pickettii 12J	—	bla_{OXA-60} 100%	no genes found
Staphylococcus aureus subsp. *aureus* COL	methicillin resistant	*aac(3)-Ik* 98.87% *mecA* 100% *tet(38)* 99.85%	*tet(K)* 100%
Streptococcus suis BM407	*tet(M)*, *tet(O)*, *tet(L)*, chloramphenicol acetyltransferase	*erm(B)* 99.86% *tet(M)* 99.83% *tet(O)* 99.64% *tet(L)* 100%	no genes found
Xanthomonas campestris pv. *vesicatoria* str. 85-10	streptomycin resistance	*strA* 100% *strB* 100%	no genes found

All sequences were run through all databases in ResFinder with a selected threshold of ID=98.00%. The following strains had no annotated resistance, and no resistance genes were detected by ResFinder: *Bacillus anthracis* str. 'Ames Ancestor', *Bacillus cereus* 03BB102, *Bacillus thuringiensis* BMB171, *Burkholderia glumae* BGR1, *Burkholderia multivorans* ATCC 17616, *Clavibacter michiganensis* subsp. *michiganensis* NCPPB 382, *Coxiella burnetii* CbuK_Q154, *Cronobacter turicensis* z3032, *Erwinia amylovora* CFBP143, *Erwinia pyrifoliae* DSM 12163, *Helicobacter pylori* B8, *Legionella longbeachae* NSW150, *Listeria monocytogenes* 08-5578, *Pantoea ananatis* AJ13355, *Ralstonia solanacearum* GMI100, *Vibrio harveyi* ATCC BAA-1116, *Vibrio vulnificus* YJ016, *Yersinia enterocolitica* subsp. *enterocolitica* 8081 and *Yersinia pseudotuberculosis* PB1+.

all four *K. pneumonia* isolates. If we had used PCR to detect genes, we would probably not have found more than one, as it is common to cease looking for more genes after a matching gene is found. ResFinder can therefore potentially give more information than the existing method.

ResFinder is a further step in our development of bioinformatics tools for analyzing WGS data; the tools are specifically designed to be easy to use—and for investigators with limited bioinformatics experience. An online tool allowing identification of multilocus sequence types is already available.[7] Additional tools under development include those for the identification of virulence genes and species, and identification and phylogenetic analysis based on single-nucleotide polymorphism and pan-genome analysis.

ResFinder will continuously be updated to include additional and novel emerging resistance genes as they are identified.

Acknowledgements

We are grateful to Inge M. Hansen and John Damm Sørensen for excellent technical assistance.

Funding

This study was supported by the Center for Genomic Epidemiology (www.genomicepidemiology.org) grant 09-067103/DSF from the Danish Council for Strategic Research and by the European Union Reference Laboratory for Antimicrobial Resistance.

Transparency declarations

None to declare.

Supplementary data

Tables S1 and S2 are available as Supplementary data at *JAC* Online (http://jac.oxfordjournals.org/).

References

1 Aarestrup FM, Agerso Y, Gerner-Smidt P *et al*. Comparison of antimicrobial resistance phenotypes and resistance genes in *Enterococcus faecalis* and *Enterococcus faecium* from humans in the community, broilers, and pigs in Denmark. *Diagn Microbiol Infect Dis* 2000; **37**: 127–37.

2 Batchelor M, Hopkins KL, Liebana E *et al*. Development of a miniaturised microarray-based assay for the rapid identification of antimicrobial resistance genes in Gram-negative bacteria. *Int J Antimicrob Agents* 2008; **31**: 440–51.

3 Hasman H, Mevius D, Veldman K *et al*. β-Lactamases among extended-spectrum β-lactamase (ESBL)-resistant *Salmonella* from poultry, poultry products and human patients in The Netherlands. *J Antimicrob Chemother* 2005; **56**: 115–21.

4 Bennedsen M, Stuer-Lauridsen B, Danielsen M *et al*. Screening for antimicrobial resistance genes and virulence factors via genome sequencing. *Appl Environ Microbiol* 2011; **77**: 2785–7.

5 Shaw KJ, Rather PN, Hare RS *et al*. Molecular genetics of aminoglycoside resistance genes and familial relationships of the aminoglycoside-modifying enzymes. *Microbiol Rev* 1993; **57**: 138–63.

6 van Hoek AH, Mevius D, Guerra B *et al*. Acquired antibiotic resistance genes: an overview. *Front Microbiol* 2011; **2**: 203.

7 Larsen MV, Consentino S, Rasmussen S *et al*. Multilocus sequence typing of total-genome-sequenced bacteria. *J Clin Microbiol* 2012; **50**: 1355–61.

8 Hendriksen RS, Seyfarth AM, Jensen AB *et al*. Results of use of WHO Global Salm-Surv external quality assurance system for antimicrobial susceptibility testing of *Salmonella* isolates from 2000 to 2007. *J Clin Microbiol* 2009; **47**: 79–85.

9 Rowe-Magnus DA, Guerout AM, Mazel D. Bacterial resistance evolution by recruitment of super-integron gene cassettes. *Mol Microbiol* 2002; **43**: 1657–69.

10 Price LB, Stegger M, Hasman H *et al*. *Staphylococcus aureus* CC398: host adaptation and emergence of methicillin resistance in livestock. *MBio* 2012; **3**: e00305–11.

5 CONCLUSIONS AND FUTURE PERSPECTIVES

In this work we show how it is possible to use bacterial WGS to characterise bacterial strains and predict pathogenicity features of potentially pathogenic strains. The web-servers developed at CGE, and discussed in this thesis, are a good example of useful tools with web-interfaces simple enough to be used by people with low computer skills and at the same time, with an high need to use bioinformatics tools for their research and work.

The steps in the near future will be the creation of databases of bacterial strains, integrated in a system in which a user can see all the tests (e.g., MLST, antibiotic resistance etc.) he has done on strains of a given bacterial species that he is working on. Such a system will allow, for example, to quickly compare a set of strains of the same species, or strains involved in the same disease, to do epidemiological studies on a given bacteria or get insight of the mechanisms behind a given infectious disease.

The next step following the research described in this thesis will be the extension of the analysis to vira and metagenomic sample. Even though there are different sets of tools for the fast analysis and characterisation of bacterial WGS data, little has been done to develop tools for a fast and comprehensive analysis of clinical metagenomic samples. A few pipelines have been developed for the phylogenetic and functional analysis of metagenomes [32, 49, 68], but there is at present no pipeline focusing on the identification of genes and groups of microbes that might be responsible for the disease.

Metagenomic data obtained from, e.g., human urine or fecal samples contain a mix of microbes and moreover frequently contain only parts of each genome, thus making the analysis much more challenging than single-culture analysis. Metagenomic analyses have, however, two advantages of pivotal importance. Firstly, single-culture analysis is more time consuming, and secondly, some disease-causing bacteria cannot be grown in culture outside their natural environment.

The analysis of metagenomic data can give us a "picture" of the groups of microbes in the host, which is wider than the information we obtain using WGS data from single bacteria. If we consider for example antibiotic resistance, analyzing metagenomic samples we can find all the antimicrobial resistant genes that are the in host and possibly associate those to specific bacteria. This could be very useful to understand, for example, how a given patient would react to a given drug before the treatment has started. This kind of assessments are done at present when a patient is admitted to a hospital in order to reduce the risk for the patient to acquire nosocomial infections due to multi-drug resistant bacteria during treatment with antimicrobial drugs. The pipeline would help to speed up this kind of assessments and, given the steadily decreasing costs for sequencing, reduce also the costs

for the healthcare related to both nosocomial infections and patients screening before treatment with antibiotics.

BIBLIOGRAPHY

[1] DANMAP | danish programme for surveillance of antimicrobial consumption and resistance: report 2012. 11

[2] New NIH awards focus on nanopore technology for DNA sequencing. 8

[3] Point prevalence survey of healthcare-associated infections and antimicrobial use in european acute care hospitals 2011-2012. 11

[4] UC davis | 100K foodborne pathogen genome project. 13

[5] WHO | the global burden of disease: 2004 update. 1

[6] WHO | the evolving threat of antimicrobial resistance - options for action, 2012. 11

[7] GR Abecasis, David Altshuler, A Auton, LD Brooks, RM Durbin, Richard A Gibbs, Matt E Hurles, Gil A McVean, DR Bentley, A Chakravarti, et al. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010. 3

[8] S Anderson. Shotgun DNA sequencing using cloned DNase i-generated fragments. *Nucleic Acids Research*, 9(13):3015–3027, July 1981. PMID: 6269069 PMCID: PMC327328. 4

[9] M. Andreatta, M. Nielsen, F.M. Aarestrup, and O. Lund. In silico prediction of human pathogenicity in the -proteobacteria. *PloS one*, 5(10):e13680, 2010. 13, 17

[10] X Autosomes Chromosome. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491:1, 2012. 4

[11] Serafim Batzoglou, David B Jaffe, Ken Stanley, Jonathan Butler, Sante Gnerre, Evan Mauceli, Bonnie Berger, Jill P Mesirov, and Eric S Lander. ARACHNE: a whole-genome shotgun assembler. *Genome research*, 12(1):177–189, 2002. 10

[12] David R. Bentley, Shankar Balasubramanian, Harold P. Swerdlow, Geoffrey P. Smith, John Milton, Clive G. Brown, Kevin P. Hall, Dirk J. Evers, Colin L. Barnes, Helen R. Bignell, Jonathan M. Boutell, Jason Bryant, Richard J. Carter, R. Keira Cheetham, Anthony J. Cox, Darren J. Ellis, Michael R. Flatbush, Niall A. Gormley, Sean J. Humphray, Leslie J. Irving, Mirian S. Karbelashvili, Scott M. Kirk, Heng Li, Xiaohai Liu, Klaus S. Maisinger, Lisa J. Murray, Bojan Obradovic, Tobias Ost, Michael L. Parkinson, Mark R. Pratt, Isabelle M. J. Rasolonjatovo, Mark T. Reed, Roberto Rigatti, Chiara Rodighiero, Mark T. Ross, Andrea Sabot, Subramanian V. Sankar, Aylwyn Scally, Gary P. Schroth, Mark E. Smith, Vincent P. Smith, Anastassia Spiridou, Peta E. Torrance, Svilen S. Tzonev, Eric H. Vermaas, Klaudia Walter, Xiaolin Wu, Lu Zhang, Mohammed D. Alam, Carole Anastasi, Ify C. Aniebo, David M. D. Bailey, Iain R. Bancarz, Saibal Banerjee, Selena G. Barbour, Primo A. Baybayan, Vincent A. Benoit, Kevin F. Benson, Claire Bevis, Phillip J. Black, Asha Boodhun, Joe S. Brennan, John A. Bridgham, Rob C. Brown, Andrew A. Brown, Dale H. Buermann, Abass A. Bundu, James C. Burrows, Nigel P. Carter, Nestor Castillo, Maria Chiara E. Catenazzi, Simon Chang, R. Neil Cooley, Natasha R. Crake, Olubunmi O. Dada, Konstantinos D. Diakoumakos, Belen Dominguez-Fernandez, David J. Earnshaw, Ugonna C. Egbujor, David W. Elmore, Sergey S. Etchin, Mark R. Ewan, Milan Fedurco, Louise J. Fraser, Karin V. Fuentes Fajardo, W. Scott Furey, David George, Kimberley J. Gietzen, Colin P. Goddard, George S. Golda, Philip A. Granieri, David E. Green, David L. Gustafson, Nancy F. Hansen, Kevin Harnish, Christian D. Haudenschild, Narinder I. Heyer, Matthew M. Hims, Johnny T. Ho, Adrian M. Horgan, Katya Hoschler, Steve Hurwitz, Denis V. Ivanov, Maria Q. Johnson, Terena James, T. A. Huw Jones, Gyoung-Dong Kang, Tzvetana H. Kerelska, Alan D. Kersey,

Irina Khrebtukova, Alex P. Kindwall, Zoya Kingsbury, Paula I. Kokko-Gonzales, Anil Kumar, Marc A. Laurent, Cynthia T. Lawley, Sarah E. Lee, Xavier Lee, Arnold K. Liao, Jennifer A. Loch, Mitch Lok, Shujun Luo, Radhika M. Mammen, John W. Martin, Patrick G. McCauley, Paul McNitt, Parul Mehta, Keith W. Moon, Joe W. Mullens, Taksina Newington, Zemin Ning, Bee Ling Ng, Sonia M. Novo, Michael J. O'Neill, Mark A. Osborne, Andrew Osnowski, Omead Ostadan, Lambros L. Paraschos, Lea Pickering, Andrew C. Pike, Alger C. Pike, D. Chris Pinkard, Daniel P. Pliskin, Joe Podhasky, Victor J. Quijano, Come Raczy, Vicki H. Rae, Stephen R. Rawlings, Ana Chiva Rodriguez, Phyllida M. Roe, John Rogers, Maria C. Rogert Bacigalupo, Nikolai Romanov, Anthony Romieu, Rithy K. Roth, Natalie J. Rourke, Silke T. Ruediger, Eli Rusman, Raquel M. Sanches-Kuiper, Martin R. Schenker, Josefina M. Seoane, Richard J. Shaw, Mitch K. Shiver, Steven W. Short, Ning L. Sizto, Johannes P. Sluis, Melanie A. Smith, Jean Ernest Sohna Sohna, Eric J. Spence, Kim Stevens, Neil Sutton, Lukasz Szajkowski, Carolyn L. Tregidgo, Gerardo Turcatti, Stephanie vandeVondele, Yuli Verhovsky, Selene M. Virk, Suzanne Wakelin, Gregory C. Walcott, Jingwen Wang, Graham J. Worsley, Juying Yan, Ling Yau, Mike Zuerlein, Jane Rogers, James C. Mullikin, Matthew E. Hurles, Nick J. McCooke, John S. West, Frank L. Oaks, Peter L. Lundberg, David Klenerman, Richard Durbin, and Anthony J. Smith. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, November 2008. 5

[13] Sébastien Boisvert, Fran\ccois Laviolette, and Jacques Corbeil. Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *Journal of Computational Biology*, 17(11):1519–1533, 2010. 10

[14] Jonathan Butler, Iain MacCallum, Michael Kleber, Ilya A Shlyakhter, Matthew K Belmonte, Eric S Lander, Chad Nusbaum, and David B Jaffe. ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome research*, 18(5):810–820, 2008. 10

[15] Benilton S. Carvalho and Gabriella Rustici. The challenges of delivering bioinformatics training in the analysis of high-throughput data. *Briefings in Bioinformatics*, page bbt018, March 2013. PMID: 23543353. 14

[16] Bastien Chevreux. MIRA: an automated genome and EST assembler. *Ruprecht-Karls University, Heidelberg, Germany*, 2005. 10

[17] Peter J. A. Cock, Christopher J. Fields, Naohisa Goto, Michael L. Heuer, and Peter M. Rice. The sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 38(6):1767–1771, April 2010. PMID: 20015970. 5

[18] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004. 2

[19] Salvatore Cosentino, Mette Voldby Larsen, Frank Møller Aarestrup, and Ole Lund. PathogenFinder - distinguishing friend from foe using bacterial whole genome sequence data. *PLoS ONE*, 8(10):e77302, October 2013. 13, 14, 15, 17

[20] Casper DJ den Heijer, Evelien ME van Bijnen, W John Paget, Mike Pringle, Herman Goossens, Cathrien A Bruggeman, François G Schellevis, and Ellen E Stobberingh. Prevalence and resistance of commensal staphylococcus aureus, including meticillin-resistant s aureus, in nine european countries: a cross-sectional study. *The Lancet Infectious Diseases*, 13(5):409–415, 2013. 11

[21] Xavier Didelot, Rory Bowden, Daniel J. Wilson, Tim E. A. Peto, and Derrick W. Crook. Transforming clinical microbiology with bacterial genome sequencing. *Nature Reviews Genetics*, 13(9):601–612, 2012. 1

[22] Devin Dressman, Hai Yan, Giovanni Traverso, Kenneth W Kinzler, and Bert Vogelstein. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proceedings of the National Academy of Sciences*, 100(15):8817–8822, 2003. 5

[23] Dent Earl, Keith Bradnam, John St John, Aaron Darling, Dawei Lin, Joseph Fass, Hung On Ken Yu, Vince Buffalo, Daniel R. Zerbino, Mark Diekhans, Ngan Nguyen, Pramila Nuwantha Ariyaratne, Wing-Kin Sung, Zemin Ning, Matthias Haimel, Jared T. Simpson, Nuno A. Fonseca, İnanç Birol, T. Roderick Docking, Isaac Y. Ho, Daniel S. Rokhsar, Rayan Chikhi, Dominique Lavenier, Guillaume Chapuis, Delphine Naquin, Nicolas Maillet, Michael C. Schatz, David R. Kelley, Adam M. Phillippy, Sergey Koren, Shiaw-Pyng Yang, Wei Wu, Wen-Chi Chou, Anuj Srivastava, Timothy I. Shaw, J. Graham Ruby, Peter Skewes-Cox, Miguel Betegon, Michelle T. Dimon, Victor Solovyev, Igor Seledtsov, Petr Kosarev, Denis Vorobyev, Ricardo Ramirez-Gonzalez, Richard Leggett, Dan MacLean, Fangfang Xia, Ruibang Luo, Zhenyu Li, Yinlong Xie, Binghang Liu, Sante Gnerre, Iain MacCallum, Dariusz Przybylski, Filipe J. Ribeiro, Shuangye Yin, Ted Sharpe, Giles Hall, Paul J. Kersey, Richard Durbin, Shaun D. Jackman, Jarrod A. Chapman, Xiaoqiu Huang, Joseph L. DeRisi, Mario Caccamo, Yingrui Li, David B. Jaffe, Richard E. Green, David Haussler, Ian Korf, and Benedict Paten. Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Research*, 21(12):2224–2241, December 2011. PMID: 21926179. 10

[24] John Eid, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, David Rank, Primo Baybayan, Brad Bettman, et al. Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910):133–138, 2009. 7

[25] Michael Eisenstein. Oxford nanopore announcement sets sequencing sector abuzz. *Nature biotechnology*, 30(4):295–296, 2012. 7

[26] R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, and Et Al. Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science*, 269(5223):496–512, July 1995. PMID: 7542800. 2

[27] Claire M. Fraser, Jeannine D. Gocayne, Owen White, Mark D. Adams, Rebecca A. Clayton, Robert D. Fleischmann, Carol J. Bult, Anthony R. Kerlavage, Granger Sutton, Jenny M. Kelley, Janice L. Fritchman, Janice F. Weidman, Keith V. Small, Mina Sandusky, Joyce Fuhrmann, David Nguyen, Teresa R. Utterback, Deborah M. Saudek, Cheryl A. Phillips, Joseph M. Merrick, Jean-Francois Tomb, Brian A. Dougherty, Kenneth F. Bott, Ping-Chuan Hu, Thomas S. Lucier, Scott N. Peterson, Hamilton O. Smith, Clyde A. Hutchison, and J. Craig Venter. The minimal gene complement of mycoplasma genitalium. *Science*, 270(5235):397–404, October 1995. PMID: 7569993. 2

[28] D N Fredericks and D A Relman. Sequence-based identification of microbial pathogens: a reconsideration of koch's postulates. *Clinical microbiology reviews*, 9(1):18–33, January 1996. PMID: 8665474. 2

[29] Sante Gnerre, Iain MacCallum, Dariusz Przybylski, Filipe J Ribeiro, Joshua N Burton, Bruce J Walker, Ted Sharpe, Giles Hall, Terrance P Shea, Sean Sykes, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences*, 108(4):1513–1518, 2011. 10

[30] J. Peter Gogarten and Jeffrey P. Townsend. Horizontal gene transfer, genome innovation and evolution. *Nature Reviews Microbiology*, 3(9):679–687, 2005. 17

[31] Robert W. Holley, Jean Apgar, George A. Everett, James T. Madison, Mark Marquisee, Susan H. Merrill, John Robert Penswick, and Ada Zamir. Structure of a ribonucleic acid. *Science*, 147(3664):1462–1465, March 1965. PMID: 14263761. 2

[32] Daniel H Huson, Alexander F Auch, Ji Qi, and Stephan C Schuster. MEGAN analysis of metagenomic data. *Genome research*, 17(3):377–386, 2007. 71

[33] Clyde A. Hutchison. DNA sequencing: bench to bedside and beyond. *Nucleic Acids Research*, 35(18):6227–6237, September 2007. PMID: 17855400. 2

[34] Gregorio Iraola, Gustavo Vazquez, Lucía Spangenberg, and Hugo Naya. Reduced set of virulence genes allows high accuracy prediction of bacterial pathogenicity in humans. *PloS one*, 7(8):e42144, 2012. 13

[35] William R Jeck, Josephine A Reinhardt, David A Baltrus, Matthew T Hickenbotham, Vincent Magrini, Elaine R Mardis, Jeffery L Dangl, and Corbin D Jones. Extending assembly of short DNA sequences to handle error. *Bioinformatics*, 23(21):2942–2944, 2007. 10

[36] Xiaoli Jiao. A benchmark study on error assessment and quality control of CCS reads derived from the PacBio RS. *Journal of Data Mining in Genomics & Proteomics*, 04(03), 2013. 7

[37] A. D. Kaiser and Ray Wu. Structure and function of DNA cohesive ends. *Cold Spring Harbor Symposia on Quantitative Biology*, 33:729–734, January 1968. PMID: 4892006. 2

[38] K. G. Kuhn, G. Sørensen, M. Torpdahl, M. K. Kjeldsen, T. Jensen, S. Gubbels, G. O. Bjerager, A. Wingstrand, L. J. Porsbo, and S. Ethelberg. A long-lasting outbreak of salmonella typhimurium u323 associated with several pork products, denmark, 2010. *Epidemiology & Infection*, 141(02):260–268, 2013. 1

[39] Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001. 2

[40] Mette V. Larsen, Salvatore Cosentino, Simon Rasmussen, Carsten Friis, Henrik Hasman, Rasmus Lykke Marvig, Lars Jelsbak, Thomas Sicheritz-Pontén, David W. Ussery, Frank M. Aarestrup, and Ole Lund. Multilocus sequence typing of total-genome-sequenced bacteria. *Journal of Clinical Microbiology*, 50(4):1355–1361, April 2012. 14, 15, 16, 31

[41] Pimlapas Leekitcharoenphon, Rolf S Kaas, Martin Christen F Thomsen, Carsten Friis, Simon Rasmussen, and Frank M Aarestrup. snpTree-a web-server to identify and construct SNP trees from whole genome sequence data. *BMC genomics*, 13(Suppl 7):S6, 2012. 14, 15, 16

[42] Ruiqiang Li, Hongmei Zhu, Jue Ruan, Wubin Qian, Xiaodong Fang, Zhongbin Shi, Yingrui Li, Shengting Li, Gao Shan, Karsten Kristiansen, et al. De novo assembly of human genomes with massively parallel short read sequencing. *Genome research*, 20(2):265–272, 2010. 10

[43] Nicholas J. Loman, Chrystala Constantinidou, Jacqueline Z. M. Chan, Mihail Halachev, Martin Sergeant, Charles W. Penn, Esther R. Robinson, and Mark J. Pallen. High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nature Reviews Microbiology*, 10(9):599–606, 2012. 6

[44] Marcel Margulies, Michael Egholm, William E. Altman, Said Attiya, Joel S. Bader, Lisa A. Bemben, Jan Berka, Michael S. Braverman, Yi-Ju Chen, Zhoutao Chen, Scott B. Dewell, Lei Du, Joseph M. Fierro, Xavier V. Gomes, Brian C. Godwin, Wen He, Scott Helgesen, Chun He Ho, Gerard P. Irzyk, Szilveszter C. Jando, Maria L. I. Alenquer, Thomas P. Jarvie, Kshama B. Jirage, Jong-Bum Kim, James R. Knight, Janna R. Lanza, John H. Leamon, Steven M. Lefkowitz, Ming Lei, Jing Li, Kenton L.

Lohman, Hong Lu, Vinod B. Makhijani, Keith E. McDade, Michael P. McKenna, Eugene W. Myers, Elizabeth Nickerson, John R. Nobile, Ramona Plant, Bernard P. Puc, Michael T. Ronan, George T. Roth, Gary J. Sarkis, Jan Fredrik Simons, John W. Simpson, Maithreyan Srinivasan, Karrie R. Tartaro, Alexander Tomasz, Kari A. Vogt, Greg A. Volkmer, Shally H. Wang, Yong Wang, Michael P. Weiner, Pengguang Yu, Richard F. Begley, and Jonathan M. Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, 2005. 5

[45] Bonnie M. Marshall and Stuart B. Levy. Food animals and antimicrobials: Impacts on human health. *Clinical Microbiology Reviews*, 24(4):718–733, October 2011. PMID: 21976606. 11, 63

[46] Vivien Marx. Biology: The big challenges of big data. *Nature*, 498(7453):255–260, June 2013. 3

[47] Michael L. Metzker. Emerging technologies in DNA sequencing. *Genome Research*, 15(12):1767–1776, December 2005. PMID: 16339375. 2

[48] Elisabeth Meyer, Petra Gastmeier, Maria Deja, and Frank Schwab. Antibiotic consumption and resistance: Data from europe and germany. *International Journal of Medical Microbiology*, 303(6-7):388–395, August 2013. 11, 63

[49] Folker Meyer, Daniel Paarmann, Mark D'Souza, Robert Olson, Elizabeth M Glass, Michael Kubal, Tobias Paczian, A Rodriguez, Rick Stevens, Andreas Wilke, et al. The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics*, 9(1):386, 2008. 71

[50] Jason R Miller, Arthur L Delcher, Sergey Koren, Eli Venter, Brian P Walenz, Anushka Brownley, Justin Johnson, Kelvin Li, Clark Mobarry, and Granger Sutton. Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics*, 24(24):2818–2824, 2008. 10

[51] Jason R. Miller, Sergey Koren, and Granger Sutton. Assembly algorithms for next-generation sequencing data. *Genomics*, 95(6):315–327, 2010. 9, 10

[52] Eugene W. Myers. Toward simplifying and accurately formulating fragment assembly. *Journal of Computational Biology*, 2(2):275–290, January 1995. 10

[53] Pavel A Pevzner and Haixu Tang. Fragment assembly with double-barreled data. *Bioinformatics*, 17(suppl 1):S225–S233, 2001. 8, 10

[54] Pavel A Pevzner, Haixu Tang, and Michael S Waterman. An eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences*, 98(17):9748–9753, 2001. 10

[55] Mihai Pop. Genome assembly reborn: recent computational challenges. *Briefings in bioinformatics*, 10(4):354–366, 2009. 10

[56] Mostafa Ronaghi, Mathias Uhlén, and P\aal Nyrén. A sequencing method based on real-time pyrophosphate. *Science*, 281(5375):363–365, 1998. 5

[57] Jonathan M Rothberg, Wolfgang Hinz, Todd M Rearick, Jonathan Schultz, William Mileski, Mel Davey, John H Leamon, Kim Johnson, Mark J Milgrew, Matthew Edwards, et al. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475(7356):348–352, 2011. 5

[58] F. Sanger. The croonian lecture, 1975: Nucleotide sequences in DNA. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 191(1104):317–333, December 1975. PMID: 2920. 2

[59] F Sanger, GM Air, BG Barrell, NL Brown, AR Coulson, JC Fiddes, PM Slocombe, and M Smith. Nucleotide sequence of bacteriophage (d x174 DNA. 1977. 2

[60] F Sanger, AR Coulson, T Friedmann, GM Air, BG Barrell, NL Brown, JC Fiddes, CA Hutchison III, PM Slocombe, and M Smith. The nucleotide sequence of bacteriophage X174. *Journal of molecular biology*, 125(2):225–246, 1978. 2

[61] Fred Sanger and Alan R Coulson. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology*, 94(3):441–448, 1975. 2

[62] Frederick Sanger, Steven Nicklen, and Alan R Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467, 1977. 2

[63] N. L. Sherry, J. L. Porter, T. Seemann, A. Watkins, T. P. Stinear, and B. P. Howden. Outbreak investigation using high-throughput genome sequencing within a diagnostic microbiology laboratory. *Journal of Clinical Microbiology*, 51(5):1396–1401, February 2013. 31

[64] R Staden. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Research*, 6(7):2601–2610, June 1979. PMID: 461197 PMCID: PMC327874. 4

[65] Bärbel Stecher, Rémy Denzler, Lisa Maier, Florian Bernet, Mandy J. Sanders, Derek J. Pickard, Manja Barthel, Astrid M. Westendorf, Karen A. Krogfelt, Alan W. Walker, Martin Ackermann, Ulrich Dobrindt, Nicholas R. Thomson, and Wolf-Dietrich Hardt. Gut inflammation can boost horizontal gene transfer between pathogenic and commensal enterobacteriaceae. *Proceedings of the National Academy of Sciences*, 109(4):1269–1274, January 2012. 17

[66] Louise H Taylor, Sophia M Latham, and EJ Mark. Risk factors for human disease emergence. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 356(1411):983–989, 2001. 11

[67] Mikkel Thorup. Near-optimal fully-dynamic graph connectivity. In *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, page 343–350, 2000. 10

[68] Todd J Treangen, Sergey Koren, Daniel D Sommer, Bo Liu, Irina Astrovskaya, Brian Ondov, Aaron E Darling, Adam M Phillippy, and Mihai Pop. MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. *Genome biology*, 14(1):R2, 2013. 71

[69] Anton Valouev, Jeffrey Ichikawa, Thaisan Tonthat, Jeremy Stuart, Swati Ranade, Heather Peckham, Kathy Zeng, Joel A Malek, Gina Costa, Kevin McKernan, et al. A high-resolution, nucleosome position map of c. elegans reveals a lack of universal sequence-dictated positioning. *Genome research*, 18(7):1051–1063, 2008. 5

[70] J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. The sequence of the human genome. *science*, 291(5507):1304–1351, 2001. 2

[71] René L Warren, Granger G Sutton, Steven JM Jones, and Robert A Holt. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics*, 23(4):500–501, 2007. 10

[72] James D Watson, Francis HC Crick, et al. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953. 2

[73] Barbara Wold and Richard M Myers. Sequence census methods for functional genomics. *Nature methods*, 5(1):19, 2008. 3

[74] M. E. J. Woolhouse and M. J. Ward. Sources of antimicrobial resistance. *Science*, 341(6153):1460–1461, September 2013. 63

[75] Ray Wu and Ellen Taylor. Nucleotide sequence analysis of DNA: II. complete nucleotide sequence of the cohesive ends of bacteriophage DNA. *Journal of molecular biology*, 57(3):491–511, 1971. 2

[76] Ea Zankari, Henrik Hasman, Salvatore Cosentino, Martin Vestergaard, Simon Rasmussen, Ole Lund, Frank M. Aarestrup, and Mette Voldby Larsen. Identification of acquired antimicrobial resistance genes. *Journal of Antimicrobial Chemotherapy*, 67(11):2640–2644, November 2012. PMID: 22782487. 14, 15, 16, 63

[77] Daniel R Zerbino and Ewan Birney. Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome research*, 18(5):821–829, 2008. 10

[78] Daniel R Zerbino, Gayle K McEwen, Elliott H Margulies, and Ewan Birney. Pebble and rock band: heuristic resolution of repeats and scaffolding in the velvet short-read de novo assembler. *PloS one*, 4(12):e8407, 2009. 10